Suplement Material: Generalization Error Bounds on Deep Learning with Markov Datasets

Lan V. Truong[*]
Department of Engineering
University of Cambridge
Cambridge, CB2 1PZ
lt407@cam.ac.uk

1 Main and Supplemental Results

1.1 Probabilistic Bounds for General Function Classes

In this section, we develop probabilistic bounds for general function classes in terms of Gaussian and Rademacher complexities. First, we prove the following key lemma, which is an extension of the symmetrization inequality for i.i.d. sequences (for example, [1]) to a new version for Markov sequences $\{X_n\}_{n=1}^{\infty}$ with the stationary distribution π and the initial distribution $\nu \in \mathcal{M}_2$:

Lemma 1. *Let $\mathcal{F}$ be a class of functions which are uniformly bounded by M. For all $n \in \mathbb{Z}^+$, define*

$$A_n := \sqrt{\frac{2M}{n(1-\lambda)} + \frac{64M^2}{n^2(1-\lambda)^2}\left\|\frac{dv}{d\pi} - 1\right\|_2}, \tag{1}$$

$$\tilde{A}_n := \frac{M}{2n}\left[\sqrt{2\tau_{\min} n \log n} + \sqrt{n} + 4\right]. \tag{2}$$

Then, the following holds:

$$\frac{1}{2}\mathbb{E}\big[\|P_n^0\|_{\mathcal{F}}\big] - \tilde{A}_n \leq \mathbb{E}\big[\big\|P_n - P\big\|_{\mathcal{F}}\big] \leq 2\mathbb{E}\big[\|P_n^0\|_{\mathcal{F}}\big] + A_n, \tag{3}$$

where

$$\|P_n^0\|_{\mathcal{F}} := \sup_{f \in \mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i f(X_i)\right|. \tag{4}$$

Proof. See Appendix A. □

Next, we prove the following theorem.

Theorem 2. *Consider a countable family of Lipschitz function $\Phi = \{\varphi_k : k \geq 1\}$, where $\varphi_k : \mathbb{R} \to \mathbb{R}$ satisfies $\mathbf{1}_{(-\infty,0]}(x) \leq \varphi_k(x)$ for all k. For each $\varphi \in \Phi$, denote by $L(\varphi)$ its Lipschitz constant. Define*

$$B_n := \sqrt{\frac{2}{n(1-\lambda)} + \frac{64}{n^2(1-\lambda)^2}\left\|\frac{dv}{d\pi} - 1\right\|_2}. \tag{5}$$

[*] Use footnote for providing further information about author (webpage, alternative address)—*not* for acknowledging funding agencies.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

Then, for any $t > 0$,

$$\mathbb{P}\bigg(\exists f \in \mathcal{F} : P\{f \le 0\} > \inf_{k>0}\bigg[P_n\varphi_k(f) + 4L(\varphi_k)R_n(\mathcal{F}) + \big(t + \sqrt{\log k}\big)\sqrt{\frac{\tau_{\min}}{n}} + B_n\bigg]\bigg)$$
$$\le \frac{\pi^2}{3}\exp(-2t^2) \tag{6}$$

and

$$\mathbb{P}\bigg(\exists f \in \mathcal{F} : P\{f \le 0\} > \inf_{k>0}\bigg[P_n\varphi_k(f) + \sqrt{2\pi}L(\varphi_k)G_n(\mathcal{F}) + \frac{2}{\sqrt{n}} + \big(t + \sqrt{\log k}\big)\sqrt{\frac{\tau_{\min}}{n}}\bigg]\bigg) \le \frac{\pi^2}{3}\exp(-2t^2). \tag{7}$$

Proof. See Appendix B. □

Theorem 3. *Let φ is a non-increasing function such that $\varphi(x) \ge \mathbf{1}_{(-\infty,0]}(x)$ for all $x \in \mathbb{R}$. For any $t > 0$,*

$$\mathbb{P}\bigg(\exists f \in \mathcal{F} : P\{f \le 0\} > \inf_{\delta\in(0,1]}\bigg[P_n\varphi\Big(\frac{f}{\delta}\Big) + \frac{8L(\varphi)}{\delta}R_n(\mathcal{F}) + \big(t + \sqrt{\log\log_2 2\delta^{-1}}\big)\sqrt{\frac{\tau_{\min}}{n}} + B_n\bigg]\bigg)$$
$$\le \frac{\pi^2}{3}\exp(-2t^2) \tag{8}$$

and

$$\mathbb{P}\bigg(\exists f \in \mathcal{F} : P\{f \le 0\} > \inf_{\delta\in(0,1]}\bigg[P_n\varphi\Big(\frac{f}{\delta}\Big) + \frac{2L(\varphi)\sqrt{2\pi}}{\delta}G_n(\mathcal{F}) + \frac{2}{\sqrt{n}} + \big(t + \sqrt{\log\log_2 2\delta^{-1}}\big)\sqrt{\frac{\tau_{\min}}{n}} + B_n\bigg]\bigg) \le \frac{\pi^2}{3}\exp(-2t^2), \tag{9}$$

where B_n is defined in (5).

Proof. See Appendix C. □

In the next statements, we use Rademacher complexities, but Gaussian complexities can be used similarly. Now, assume that φ is a function from $\mathbb{R}$ to $\mathbb{R}$ such that $\varphi(x) \le I_{(-\infty,0]}(x)$ for all $x \in \mathbb{R}$ and φ satisfies the Lipschitz with constant $L(\varphi)$. Then, the following theorems can be proved by using similar arguments as the proof of Theorem 3.

Theorem 4. *Let φ is a nonincreasing function such that $\varphi(x) \le \mathbf{1}_{(-\infty,0]}(x)$ for all $x \in \mathbb{R}$. For any $t > 0$,*

$$\mathbb{P}\bigg(\exists f \in \mathcal{F} : P\{f \le 0\} < \sup_{\delta\in(0,1]}\bigg[P_n\varphi\Big(\frac{f}{\delta}\Big) + \frac{8L(\varphi)}{\delta}R_n(\mathcal{F}) + \big(t + \sqrt{\log\log_2 2\delta^{-1}}\big)\sqrt{\frac{\tau_{\min}}{n}} + B_n\bigg]\bigg) \le \frac{\pi^2}{3}\exp(-2t^2) \tag{10}$$

and

$$\mathbb{P}\bigg(\exists f \in \mathcal{F} : P\{f \le 0\} < \sup_{\delta\in(0,1]}\bigg[P_n\varphi\Big(\frac{f}{\delta}\Big) + \frac{2\sqrt{2\pi}L(\varphi)}{\delta}G_n(\mathcal{F}) + \frac{2}{\sqrt{n}}\bigg]\bigg) + \big(t + \sqrt{\log\log_2 2\delta^{-1}}\big)\sqrt{\frac{\tau_{\min}}{n}} + B_n\bigg]\bigg) \le \frac{\pi^2}{3}\exp(-2t^2), \tag{11}$$

where B_n is defined in (5).

By combining Theorem 3 and Theorem 4, we obtain the following result.

Theorem 5. *Let*

$$\Delta_n(\mathcal{F};\delta) := \frac{8}{\delta} R_n(\mathcal{F}) + \sqrt{\frac{\tau_{\min} \log\log_2 2\delta^{-1}}{n}} + B_n. \tag{12}$$

Then, for all $t > 0$*,*

$$\mathbb{P}\bigg(\exists f \in \mathcal{F} : \big|P_n\{f \le 0\} - P\{f \le 0\}\big| > \inf_{\delta\in(0,1]} \bigg(P_n\{|f| \le \delta\} + \Delta_n(\mathcal{F};\delta) + t\sqrt{\frac{\tau_{\min}}{n}}\bigg)\bigg) \\ \le \frac{2\pi^2}{3}\exp(-2t^2) \tag{13}$$

and

$$\mathbb{P}\bigg(\exists f \in \mathcal{F} : \big|P_n\{f \le 0\} - P\{f \le 0\}\big| > \inf_{\delta\in(0,1]} \bigg(P\{|f| \le \delta\} + \Delta_n(\mathcal{F};\delta) + t\sqrt{\frac{\tau_{\min}}{n}}\bigg) \\ \le \frac{2\pi^2}{3}\exp(-2t^2). \tag{14}$$

Proof. Equation (13) is drawn by setting $\varphi(x) = \mathbf{1}\{x \le 0\} + (1-x)\mathbf{1}\{0 \le x \le 1\}$ in Theorem 3 and Theorem 4. Equation (14) is drawn by setting $\varphi(x) = \mathbf{1}\{x \le -1\} - x\mathbf{1}\{-1 \le x \le 0\}$ in these theorems. □

1.2 Conditions on Random Entropies and γ-Margins

As [2], given a metric space (T, d), we denote by $H_d(T;\varepsilon)$ the ε-entropy of T with respect to d, that is

$$H_d(T;\varepsilon) := \log N_d(T;\varepsilon), \tag{15}$$

where $N_d(T;\varepsilon)$ is the minimal number of balls of radius ε covering T. Let $d_{P_n,2}$ denote the metric of the space $L_2(\mathcal{S}; dP_n)$:

$$d_{P_n,2}(f,g) := \big(P_n|f-g|^2\big)^{1/2}. \tag{16}$$

For each $\gamma \in (0,1]$, define

$$\delta_n(\gamma; f) := \sup\left\{\delta \in (0,1) : \delta^{\frac{\gamma}{2}} P(f \le \delta) \le n^{-\frac{1}{2}+\frac{\gamma}{4}}\right\} \tag{17}$$

and

$$\hat{\delta}_n(\gamma; f) := \sup\left\{\delta \in (0,1) : \delta^{\frac{\gamma}{2}} P_n(f \le \delta) \le n^{-\frac{1}{2}+\frac{\gamma}{4}}\right\}. \tag{18}$$

We call $\delta_n(\gamma; f)$ and $\hat{\delta}_n(\gamma; f)$, respectively, the γ-*margin* and *empirical* γ-*margin* of f.

Theorem 6. *Suppose that for some* $\alpha \in (0,2)$ *and some constant* $D > 0$*,*

$$H_{d_{P_n},2}(\mathcal{F}; u) \le D u^{-\alpha}, \qquad u > 0 \qquad a.s., \tag{19}$$

Then, for any $\gamma \ge \frac{2\alpha}{2+\alpha}$*, there exists some constants* $\zeta, \upsilon > 0$ *such that when* n *is large enough,*

$$\mathbb{P}\bigg[\forall f \in \mathcal{F} : \zeta^{-1}\hat{\delta}_n(\gamma; f) \le \delta_n(\gamma; f) \le \zeta\hat{\delta}_n(\gamma; f)\bigg] \ge 1 - 2\upsilon\big(\log_2\log_2 n\big)\exp\big\{-n^{\frac{\gamma}{2}}/2\big\}. \tag{20}$$

Proof. See Appendix D for a detailed proof. □

1.3 Convergence rates of empirical margin distributions

First, we prove the following lemmas.

Lemma 7. *For any class $\mathcal{F}$ of bounded measurable functions from $\mathcal{S} \to \mathbb{R}$, with probability at least $1 - 2\exp\big(-2t^2\big)$, the following holds:*

$$\sup_{f\in\mathcal{F}}\sup_{y\in\mathbb{R}}\big|P_n(f \leq y) - P(f \leq y)\big| \leq \sqrt{B_n} + t\sqrt{\frac{\tau_{\min}}{n}}, \tag{21}$$

where B_n is defined in (5).

Remark 8. *By setting $t = \sqrt{2\log n}$,* (21) *shows that* $\sup_{f\in\mathcal{F}}\sup_{y\in\mathbb{R}}\big|P_n(f \leq y) - P(f \leq y)\big| \to 0$ *as $n \to \infty$.*

Proof. See Appendix E. □

Now, for each $f \in \mathcal{F}$, define

$$F_f(y) := P\{f \leq y\}, \qquad F_{n,f} := P_n\{f \leq y\}, \qquad y \in \mathbb{R}. \tag{22}$$

Let L denote the Lévy distance between the distributions in $\mathbb{R}$:

$$L(F,G) := \inf\{\delta > 0 : F(t) \leq G(t+\delta) + \delta, \qquad G(t) \leq F(t+\delta) + \delta, \qquad \forall t \in \mathbb{R}\}. \tag{23}$$

Lemma 9. *Let $M > 0$ and $\mathcal{F}$ be a class of measurable functions from $\mathcal{S}$ into $[-M, M]$. Let φ be equal to 1 for $x \leq 0$, 0 for $x \geq 1$ and linear between them. Define*

$$\tilde{\mathcal{G}}_\varphi := \left\{\varphi \circ \left(\frac{f-y}{\delta}\right) - 1 : f \in \mathcal{F}, \quad y \in [-M, M]\right\} \tag{24}$$

for some $\delta > 0$. Recall the definition of B_n in (5)*. Then, for all $t > 0$ and $\delta > 0$, the following holds:*

$$\mathbb{P}\left\{\sup_{f\in\mathcal{F}} L(F_f, F_{f,n}) \geq \delta + \mathbb{E}\big[\|P_n^0\|_{\tilde{\mathcal{G}}_\varphi}\big] + B_n + t\sqrt{\frac{\tau_{\min}}{n}}\right\} \leq 2\exp(-2t^2). \tag{25}$$

Especially, for all $t > 0$, we have

$$\mathbb{P}\left\{\sup_{f\in\mathcal{F}} L(F_f, F_{f,n}) \geq 4\sqrt{\mathbb{E}[\|P_n^0\|_{\mathcal{F}}] + M/\sqrt{n}} + B_n + t\sqrt{\frac{\tau_{\min}}{n}}\right\} \leq 2\exp(-2t^2). \tag{26}$$

Proof. See Appendix F. □

In what follows, for a function f from $\mathcal{S}$ into $\mathbb{R}$ and $M > 0$, we denote by f_M the function that is equal to f if $|f| \leq M$, is equal to M if $f > M$ and is equal to $-M$ if $f < -M$. We set

$$\mathcal{F}_M := \big\{f_M : f \in \mathcal{F}\big\}. \tag{27}$$

As always, a function $\mathcal{F}$ from $\mathcal{S}$ into $[0, \infty)$ is called an envelope of $\mathcal{F}$ iff $|f(x)| \leq F(x)$ for all $f \in \mathcal{F}$ and all $x \in \mathcal{S}$.

We write $\mathcal{F} \in \mathrm{GC(P)}$ iff $\mathcal{F}$ is a Glivenko-Cantelli class with respect to P (i.e., $\|P_n - P\|_{\mathcal{F}} \to 0$ as $n \to \infty$ a.s.). We write $\mathcal{F} \in \mathrm{BCLT(P)}$ and say that $\mathcal{F}$ satisfies the Bounded Central Limit Theorem for P iff

$$\mathbb{E}\big[\big\|P_n - P\big\|_{\mathcal{F}}\big] = O(n^{-1/2}). \tag{28}$$

Based on Lemma and Lemma 9, we prove the following theorems.

Theorem 10. *Suppose that*

$$\sup_{f\in\mathcal{F}} P\{|f| \geq M\} \to 0 \qquad as \qquad M \to \infty. \tag{29}$$

Then, the following two statements are equivalent:

- *(i)* $\mathcal{F}_M \in \mathrm{GC}(\mathrm{P})$ *for all* $M > 0$
  and

- *(ii)* $\sup_{f\in\mathcal{F}} L(F_{n,f}, F_f) \to 0 \quad$ *a.s.* $\quad n \to \infty.$

Proof. See Appendix G. □

Next, the following theorems hold.

Theorem 11. *[2, Theorem 7] The following two statements are equivalent:*

- *(i)* $\mathcal{F} \in \mathrm{GC}(\mathrm{P})$ *for all* $M > 0$
  and

- *(ii) there exists a P-integrable envelope for the class $\mathcal{F}^{(c)} = \{f - Pf : f \in \mathcal{F}\}$ and* $\sup_{f\in\mathcal{F}} L(F_{n,f}, F_f) \to 0 \qquad n \to \infty$ *and*

$$\sup_{f\in\mathcal{F}} L(F_{n,f}, F_f) \to 0 \qquad a.s. \qquad n \to \infty. \tag{30}$$

Now, we prove the following theorem.

Theorem 12. *Suppose that the class $\mathcal{F}$ is uniformly bounded. If $\mathcal{F} \in \mathrm{BCLT}(\mathrm{P})$, then*

$$\sup_{f\in\mathcal{F}} L(F_{n,f}, F_f) = O_P\biggl(\Bigl(\frac{\log n}{n}\Bigr)^{1/4}\biggr) \qquad n \to \infty. \tag{31}$$

Moreover, for some $\alpha \in (0,2)$ and for some $D > 0$

$$H_{d_{P_n,2}}(\mathcal{F}; u) \le D u^{-\alpha} \log n, \qquad u > 0, \qquad a.s., \tag{32}$$

then

$$\sup_{f\in\mathcal{F}} L(F_{n,f}, F_f) = O\bigl(n^{-\frac{1}{2+\alpha}} \log n\bigr) \qquad n \to \infty, \qquad a.s., \tag{33}$$

Proof. Appendix H. □

1.4 Bounding the generalization error of convex combinations of classifiers

We start with an application of the inequalities in Subsection 1.1 to bounding the generalization error in general classification problems. Assume that the labels take values in a finite set $\mathcal{Y}$ with $|\mathcal{Y}| = K$. Consider a class $\tilde{\mathcal{F}}$ of functions from $\mathcal{S} \times \mathcal{Y}$ into $\mathbb{R}$. A function $f \in \tilde{\mathcal{F}}$ predicts a label $y \in \mathcal{Y}$ for an example $x \in \mathcal{S}$ iff

$$f(x, y) > \max_{y' \neq y} f(x, y'). \tag{34}$$

In practice, $f(x, y)$ can be set equal to $P(y|x)$, so $\mathcal{F}$ can be assumed to be uniformly bounded. The margin of a labelled example (x, y) is defined as

$$m_{f,y}(x) := f(x, y) - \max_{y' \neq y} f(x, y'), \tag{35}$$

so f mis-classifies the label example (x, y) if and only if $m_{f,y} \le 0$. Let

$$\mathcal{F} := \bigl\{f(\cdot, y) : y \in \mathcal{Y}, f \in \tilde{\mathcal{F}}\bigr\}. \tag{36}$$

Then, we can show the following theorem.

Theorem 13. *For all $t > 0$, it holds that*

$$\mathbb{P}\biggl(\exists f \in \mathcal{F} : P\{m_{f,y} \le 0\} > \inf_{\delta\in(0,1]} \biggl[P_n\{m_{f,y} \le \delta\} + \frac{8}{\delta}(2K-1)R_n(\mathcal{F}) + \bigl(t + \sqrt{\log\log_2 2\delta^{-1}}\bigr)\sqrt{\frac{\tau_{\min}}{n} + B_n}\biggr]\biggr) \le \frac{\pi^2}{3}\exp(-2t^2), \tag{37}$$

where B_n is defined in (5)*. Here,*

$$P\{m_{f,y} \le 0\} := \sum_{x \in \mathcal{S}} \pi(x) \mathbf{1}\{m_{f,y}(x) \le 0\}, \tag{38}$$

and

$$P_n\{m_{f,y} \le \delta\} = \frac{1}{n} \sum_{k=1}^{n} \mathbf{1}\{m_{f,y}(X_k) \le \delta\} \tag{39}$$

is the empirical distribution of the Markov process $\{m_{f,y}(X_n)\}_{n=1}^{\infty}$ given f.

Proof. First, we need to bound the Rademacher's complexity for the class of functions $\big\{m_{f,y} : f \in \tilde{\mathcal{F}}\big\}$. Observe that

$$\mathbb{E}\bigg[\sup_{f \in \tilde{\mathcal{F}}} \bigg| n^{-1} \sum_{j=1}^{n} \varepsilon_j m_{f,y}(X_j)\bigg|\bigg]. \tag{40}$$

By [2, Proof of Theorem 11], we have

$$\mathbb{E}\bigg[\sup_{f \in \tilde{\mathcal{F}}} \bigg| n^{-1} \sum_{j=1}^{n} \varepsilon_j m_{f,y}(X_j)\bigg|\bigg] \le (2K-1) R_n(\mathcal{F}), \tag{41}$$

where $R_n(\mathcal{F})$ is *the Rademacher complexity function* of the class $\mathcal{F}$. Now, assume that this class of function is uniformly bounded as in practice. Hence, by Theorem 3 for φ that is equal to 1 on $(-\infty, 0]$, is equal to 0 on $[1, +\infty)$ and is linear in between, we obtain (37). □

In addition, by using the fact that $(X_1, Y_1) - (X_2, Y_2) \cdots - (X_n, Y_n)$ forms a Markov chain with stationary distribution $\tilde{\pi}$ (see the discussion on Section 2.2 in the main document), by applying Theorem 3, we obtain the following result:

Theorem 14. *Let φ is a nonincreasing function such that $\varphi(x) \ge \mathbf{1}_{(-\infty,0]}(x)$ for all $x \in \mathbb{R}$. For any $t > 0$,*

$$\begin{aligned}\mathbb{P}\bigg(\exists f \in \tilde{\mathcal{F}} : P\{\tilde{f} \le 0\} > \inf_{\delta \in (0,1]} \bigg[P_n \varphi\bigg(\frac{\tilde{f}}{\delta}\bigg) + \frac{8L(\varphi)}{\delta} R_n(\mathcal{H}) \\ + \big(t + \sqrt{\log \log_2 2\delta^{-1}}\big)\sqrt{\frac{\tau_{\min}}{n}} + B_n\bigg]\bigg) \le \frac{\pi^2}{3} \exp(-2t^2),\end{aligned} \tag{42}$$

where B_n is defined in (5)*.*

As in [2], in the voting methods of combining classifiers, a classifier produced at each iteration is a convex combination $\tilde{f}$ of simple base classifiers from the class $\mathcal{H}$. In addition, the Rademacher complexity can be bounded above by

$$R_n(\mathcal{H}) \le C \sqrt{\frac{V(\mathcal{H})}{n}} \tag{43}$$

for some constant $C > 0$, where $V(\mathcal{H})$ is the VC-dimesion of $\mathcal{H}$. Let φ be equal to 1 on $(-\infty, 0]$, is equal to 0 on $[1, +\infty)$ and is linear in between. By setting $t_\alpha = \sqrt{\frac{1}{2} \log \frac{\pi^2}{3\alpha}}$, from Theorem 14, with probability at least $1 - \alpha$, it holds that

$$P\{\tilde{f} \le 0\} \le \inf_{\delta \in (0,1]} \bigg[P_n\{\tilde{f} \le \delta\} + \frac{8C}{\delta}\sqrt{\frac{V(\mathcal{H})}{n}} + \big(t_\alpha + \sqrt{\log \log_2 2\delta^{-1}}\big)\sqrt{\frac{\tau_{\min}}{n}} + B_n\bigg], \tag{44}$$

which extends the result of Bartlett et al. [4] to Markov dataset (PAC-bound).

1.5 Bounding the generalization error in neural network learning

In this section, we consider the same example as [2, Section 6]. However, we assume that feature vectors in this dataset is generated by a Markov chain instead of an i.i.d. process. Let $\mathcal{H}$ be a class of measurable functions from $\mathcal{S} \to \mathbb{R}$ (base functions). Let $\tilde{\mathcal{H}}$ be the set of function $\tilde{f} : \mathcal{S} \times \mathcal{Y} \to \mathbb{R}$. The introduction of $\tilde{\mathcal{H}}$ is to deal with the new Markov chain $\{(X_n, Y_n)\}_{n=1}^{\infty}$ which is generated by both feature vectors and their labels instead of the feature-based Markov chain $\{X_n\}_{n=1}^{\infty}$.

Consider a feed-forward neural network where the set V of all the neurons is divided into layers

$$V = \{v_i\} \cup \bigcup_{j=0}^{l} V_j \tag{45}$$

where $V_l = \{v_o\}$. The neurons v_i and v_o are called the input and the output neurons, respectively. To define the network, we will assign the labels to the neurons in the following way. Each of the base neurons is labelled by a function from the base class $\mathcal{H}$. Each neuron of the j-th layer V_j, where $j \geq 1$, is labelled by a vector $w := (w_1, w_2, \cdots, w_m) \in \mathbb{R}^m$, where m is the number of inputs of the neuron, i.e. $m = |\mathcal{H}| + 1$. Here, w will be called the vector of weights of the neuron.

Given a Borel function σ from $\mathbb{R}$ into $[-1, 1]$ (for example, sigmoid function) and a vector $w := (w_1, w_2, \cdots, w_m) \in \mathbb{R}^m$, let

$$N_{\sigma,w} : \mathbb{R}^m \to \mathbb{R}, \qquad N_{\sigma,w}(u_1, u_2, \cdots, u_m) := \sigma\biggl(\sum_{j=1}^{m} w_j u_j\biggr). \tag{46}$$

For $w \in \mathbb{R}^m$,

$$\|w\|_1 := \sum_{i=1}^{m} |w_i|. \tag{47}$$

Let $\sigma_j : j \geq 1$ be functions from $\mathbb{R}$ into $[-1, 1]$, satisfying the Lipschitz conditions

$$\bigl|\sigma_j(u) - \sigma_j(v)\bigr| \leq L_j |u - v|, \qquad u, v \in \mathbb{R}. \tag{48}$$

The network works the following way. The input neuron inputs an instance $x \in \mathcal{S}$. A base neuron computes the value of the base function on this instance and outputs the value through its output edges. A neuron in the j-th layer ($j \geq 1$) computes and outputs through its output edges the value $N_{\sigma_j,w}(u_1, u_2, \cdots, u_m)$ (where $u_1, u_2, \cdots, u_m$ are the values of the inputs of the neuron). The network outputs the value $f(x)$ (of a function f it computes) through the output edge.

We denote by $\mathcal{N}_l$ the set of such networks. We call $\mathcal{N}_l$ the class of feed-forward neural networks with base $\mathcal{H}$ and l layers of neurons (and with sigmoid σ_j). Let $\mathcal{N}_\infty := \bigcup_{j=1}^{\infty} \mathcal{N}_j$. Define $\mathcal{H}_0 := \mathcal{H}$, and then recursively

$$\mathcal{H}_j := \biggl\{N_{\sigma_j,w}(h_1, h_2, \cdots, h_m) : m \geq 0, h_i \in \mathcal{H}_{j-1}, w \in \mathbb{R}^m\biggr\} \cup \mathcal{H}_{j-1}. \tag{49}$$

Denote $\mathcal{H}_\infty := \bigcup_{j=1}^{\infty} \mathcal{H}_j$. Clearly, $\mathcal{H}_\infty$ includes all the functions computable by feed-forward neural networks with base $\mathcal{H}$.

Let $\{b_j\}$ be a sequence of positive numbers. We also define recursively classes of functions computable by feed-forward neural networks with restrictions on the weights of neurons:

$$\mathcal{H}_j(b_1, b_2, \cdots, b_j) \tag{50}$$

$$:= \biggl\{N_{\sigma_j,w}(h_1, h_2, \cdots, h_m) : n \geq 0, h_i \in \mathcal{H}_{j-1}(b_1, b_2, \cdots, b_{j-1}), w \in \mathbb{R}^m, \|w\|_1 \leq b_j\biggr\}$$

$$\bigcup \mathcal{H}_{j-1}(b_1, b_2, \cdots, b_{j-1}). \tag{51}$$

Clearly,

$$\mathcal{H}_\infty = \bigcup \biggl\{\mathcal{H}_j(b_1, b_2, \cdots, b_j) : b_1, b_2, \cdots, b_j < +\infty\biggr\}. \tag{52}$$

As in the previous section, let φ be a function such that $\varphi(x) \geq I_{(-\infty,0]}$ for all $x \in \mathbb{R}$ and φ satisfies the Lipschitz condition with constant $L(\varphi)$. Then, the following is a direct application of Theorem 3.

Theorem 15. *For any $t \geq 0$ and for all $l \geq 1$,*

$$\mathbb{P}\Bigg(\exists f \in \mathcal{H}(b_1, b_2, \cdots, b_l) : P\{\tilde{f} \leq 0\} \\ > \inf_{\delta \in (0,1]} \Bigg[P_n \varphi\Big(\frac{\tilde{f}}{\delta}\Big) + \frac{2\sqrt{2\pi}L(\varphi)}{\delta} \prod_{j=1}^{l} (2L_j b_j + 1) G_n(\mathcal{H}) \\ + \frac{2}{\sqrt{n}} + \big(t + \sqrt{\log\log_2 2\delta^{-1}}\big)\sqrt{\frac{\tau_{\min}}{n}} + B_n \Bigg] \Bigg) \leq \frac{\pi^2}{3} \exp\big(-2t^2\big), \tag{53}$$

where B_n is defined in (5).

Proof. Let

$$\mathcal{H}'_l := \mathcal{H}(b_1, b_2, \cdots, b_l). \tag{54}$$

As the proof of [2, Theorem 13], it holds that

$$G_n(\mathcal{H}'_l) := \mathbb{E}\Big\| n^{-1} \sum_{i=1}^{n} g_i \delta_{X_i} \Big\|_{\mathcal{H}'_l} \leq \prod_{j=1}^{l} (2L_j b_j + 1) G_n(\mathcal{H}). \tag{55}$$

Hence, (53) is a direct application of Theorem 3 and (55). □

Now, given a neural network $f \in \mathcal{N}_\infty$, let

$$l(f) := \min\{j \geq 1 : f \in \mathcal{N}_j\}. \tag{56}$$

For a number k, $1 \leq k \leq l(f)$, let $V_k(f)$ denote the set of all neurons of layer k in the graph representing f. Denote

$$W_k(f) := \max_{u \in V_k(f)} \|w^{(u)}\|_1 \vee b_k, \qquad k = 1, 2, \cdots, l(f) \tag{57}$$

where $w^{(u)}$ is the coefficient vector associated with the neuron u of the layer k, and let

$$\Lambda(f) := \prod_{k=1}^{l(f)} (4L_k W_k(f) + 1), \tag{58}$$

$$\Gamma_\alpha(f) := \sum_{k=1}^{l(f)} \sqrt{\frac{\alpha}{2} \log(2 + \log_2 W_k(f))}, \tag{59}$$

where $\alpha > 0$ is the number such that $\zeta(\alpha) < 3/2$, ζ being the Riemann zeta-function:

$$\zeta(\alpha) := \sum_{k=1}^{\infty} k^{-\alpha}. \tag{60}$$

Then, by using the same arguments as [2, Proof of Theorem 14], we obtain the following result.

Theorem 16. *For any $t \geq 0$ and for all $l \geq 1$,*

$$\mathbb{P}\Bigg(\exists f \in \mathcal{H}_\infty : P\{\tilde{f} \leq 0\} > \inf_{\delta \in (0,1]} \Bigg[P_n \varphi\Big(\frac{\tilde{f}}{\delta}\Big) + \frac{2\sqrt{2\pi}L(\varphi)}{\delta} \Lambda(f) G_n(\mathcal{H}) \\ + \frac{2}{\sqrt{n}} + \big(t + \Gamma_\alpha(f) + \sqrt{\log\log_2 2\delta^{-1}}\big)\sqrt{\frac{\tau_{\min}}{n}} + B_n \Bigg] \Bigg) \\ \leq \frac{\pi^2}{3} \big(3 - 2\zeta(\alpha)\big)^{-1} \exp\big(-2t^2\big), \tag{61}$$

where B_n is defined in (5).

Proof. Let

$$\Delta_k := \begin{cases} [2^{k-1}, 2^k) & \text{for} \quad k \in \mathbb{Z}, k \neq 0, 1 \\ [1/2, 2) & \text{for} k = 1. \end{cases} \tag{62}$$

Then, using the following partition:

$$\{f \in \mathcal{H}_\infty\} = \bigcup_{l=0}^{\infty} \bigcup_{k_1 \in \mathbb{Z}\setminus\{0\}} \bigcup_{k_2 \in \mathbb{Z}\setminus\{0\}} \cdots \bigcup_{k_l \in \mathbb{Z}\setminus\{0\}} \Big\{ f \in \mathcal{H}_\infty : l(f) = l, W_j(f) \in \Delta_{k_j}, \ \forall j \in [l] \Big\}. \tag{63}$$

On each subset $\Big\{ f \in \mathcal{H}_\infty : l(f) = l, W_j(f) \in \Delta_{k_j}, \ \forall j \in [l] \Big\}$, we can lower bound $\Lambda(f)$ and $\Gamma_\alpha(f)$ by

$$\Lambda(f) \geq \prod_{j=1}^{l} (2L_j 2^{k_j} + 1), \tag{64}$$

$$\Gamma_\alpha(f) \geq \sum_{j=1}^{l} \sqrt{\frac{\alpha}{2} \log(|k_j| + 1)}. \tag{65}$$

By replacing t by $t + \sum_{j=1}^{l} \sqrt{\frac{\alpha}{2}\log(k_j+1)}$ and using Theorem 3 to bound the probability of each event and then using the union bound, we can show that

$$\begin{aligned}
&\mathbb{P}\bigg(\exists f \in \mathcal{H}_\infty : P\{\tilde{f} \leq 0\} > \inf_{\delta \in (0,1]} \bigg[P_n \varphi\bigg(\frac{\tilde{f}}{\delta}\bigg) + \frac{2\sqrt{2\pi} L(\varphi)}{\delta} \Lambda(f) G_n(\mathcal{H}) \\
&\qquad + \frac{2}{\sqrt{n}} + \big(t + \Gamma_\alpha(f) + \sqrt{\log\log_2 2\delta^{-1}}\big) \sqrt{\frac{\tau_{\min}}{n}} + B_n \bigg] \bigg) &(66)\\
&\leq \mathbb{P}\bigg(\exists f \in \mathcal{H}_\infty : P\{\tilde{f} \leq 0\} \\
&\qquad > \inf_{\delta \in (0,1]} \bigg[P_n \varphi\bigg(\frac{\tilde{f}}{\delta}\bigg) + \frac{2\sqrt{2\pi} L(\varphi)}{\delta} \Lambda(f) G_n(\mathcal{H}) \\
&\qquad\quad + \bigg(t + \sum_{j=1}^{l} \sqrt{\frac{\alpha}{2} \log(k_j + 1)} + \sqrt{\log\log_2 2\delta^{-1}} \bigg) \sqrt{\frac{\tau_{\min}}{n}} + \frac{2}{\sqrt{n}} + B_n \bigg] \bigg) &(67)\\
&\leq \sum_{l=0}^{\infty} \sum_{k_1 \in \mathbb{Z}\setminus\{0\}} \cdots \sum_{k_l \in \mathbb{Z}\setminus\{0\}} 2 \exp\bigg(-2 \bigg(t + \sum_{j=1}^{l} \sqrt{\frac{\alpha}{2} \log(k_j + 1)} \bigg)^2 \bigg) &(68)\\
&\leq \frac{\pi^2}{3} \big(3 - 2\zeta(\alpha)\big)^{-1} \exp\big(-2t^2\big), &(69)
\end{aligned}$$

where the last equation is followed by using some algebraic manipulations. □

2 Extension to High-Order Markov Chains

In this section, we extend our results in previous sections to m-order Markov chains and a mixture of m independent Markov services.

2.1 Extend to m-order Markov chain

In this subsection, we extend our results in previous sections to m-order homogeneous Markov chain. The main idea is to convert m-order homogeneous Markov chains to 1-order homogeneous Markov chain and use our results in previous sections to bound the generalization error.

We start with the following simple example.

Example 17. *[m-order moving average process without noise] Consider the following m-order Markov chain*

$$X_k = \sum_{i=1}^{m} a_i X_{k-i}, \qquad k \in \mathbb{Z}_+. \tag{70}$$

Let $Y_k := [X_{k+m-1}, X_{k+m-2}, \cdots, X_k]^T$. Then, from (70)*, we obtain*

$$Y_{k+1} = \mathbf{G} Y_k, \qquad \forall k \in \mathbb{Z}_+ \tag{71}$$

where

$$\mathbf{G} := \begin{pmatrix} a_1 & a_2 & \cdots & a_{m-1} & a_m \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{pmatrix}. \tag{72}$$

It is clear that $\{\mathbf{Y}_n\}_{n=1}^{\infty}$ is an order-1 Markov chain. Hence, instead of directly working with the m-order Markov chain $\{\mathbf{X}_n\}_{n=1}^{\infty}$, we can find an upper bound for the Markov chain $\{Y_n\}_{n=1}^{\infty}$.

To derive generalization error bounds for the Markov chain $\{Y_n\}_{n=1}^{\infty}$, we can use the following arguments. For all $f \in \mathcal{F}$ and $(x_k, x_{k+1}, \cdots, x_{k+m-1}) \in \mathcal{S}^m$, by setting $\tilde{f}(x_k, x_{k+1}, \cdots, x_{k+m-1}) = f(x_k)$ where $\tilde{f} : \mathcal{S}^m \to \mathbb{R}$, we obtain

$$\frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{f(X_i) \leq 0\} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{\tilde{f}(Y_i) \leq 0\}. \tag{73}$$

Hence, by applying all the results for 1-order Markov chain $\{Y_n\}_{n=1}^{\infty}$, we obtain corresponding upper bounds for the sequence of m-order Markov chain $\{X_n\}_{n=1}^{\infty}$.

This approach can be extended to more general m-order Markov chain $X_k = g(X_{k-1}, X_{k-2} \cdots, X_{k-m})$ where $g : \mathcal{S}^m \to \mathbb{R}$. More specifically, for any tuple $(x_1, x_2, \cdots, x_m) \in \mathcal{S}^m$, observe that

$$dg = \frac{\partial g}{\partial x_1} dx_1 + \frac{\partial g}{\partial x_2} dx_2 + \cdots + \frac{\partial g}{\partial x_m} dx_m. \tag{74}$$

Hence, if $\frac{\partial g}{\partial x_i} = \alpha_i$ for some constant α_i and for each $i \in [m]$, from (74), we have

$$g(x_1, x_2, \cdots, x_m) = g(c_1, c_2, \cdots, c_m) + \sum_{i=1}^{m} a_i x_i + \sum_{i=1}^{m} a_i \nu_i, \tag{75}$$

where ν_i's are constants. One specific example where the function $g : \mathcal{S}^m \to \mathbb{R}$ satisfies this property is $g(x_1, x_2, \cdots, x_m) = a_1 x_1 + a_2 x_2 + \cdots a_m x_m$, where $a_1, a_2, \cdots, a_m$ are constants as in Example 17.

Now, by choosing $u = \big(g(c_1, c_2, \cdots, c_m) + \sum_{i=1}^{m} \alpha_i \nu_i\big) / (1 - \sum_{i=1}^{m} a_i \nu_i)$, from (75), we have

$$g(x_1, x_2, \cdots, x_m) + u = \sum_{i=1}^{m} a_i \big(x_i + u\big). \tag{76}$$

By setting $Y_k = \big[X_{k+m-1} + u \quad X_{k+m-2} + u \quad \cdots \quad X_k + u\big]^T$, from (76), we have:

$$Y_{k+1} = \begin{pmatrix} a_1 & a_2 & \cdots & a_{m-1} & a_m \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{pmatrix} Y_k. \tag{77}$$

In a more general setting, if $X_k = g(X_{k-1}, \cdots, X_{k-m}, V_k)$ for some random variable V_k which is independent of $\{X_{k-i}\}_{i=1}^{m}$ such as the Autoregressive model (AR), where

$$X_k = c + \sum_{i=1}^{m} a_i X_{t-i} + V_k, \tag{78}$$

we can use the following conversion procedure. First, by using Taylor's approximation (to the first-order), we obtain

$$\begin{aligned} g(x_1, x_2, \cdots, x_m, \xi) \approx g(c_1, c_2, \cdots, c_m, \xi_0) + \sum_{i=1}^{m} \frac{\partial g}{\partial x_i}\bigg|_{(c_1, c_2, \cdots, c_m, \zeta_0)} (x_i - c_i) \\ + \frac{\partial g}{\partial v}\bigg|_{(c_1, c_2, \cdots, c_m, \xi_0)} (\xi - \xi_0) \end{aligned} \tag{79}$$

for some good choice of $(c_1, c_2, \cdots, c_m, \xi_0) \in \mathcal{S}^m \times \mathcal{V}$, where $\mathcal{V}$ is the alphabet of V_k. Using the above trick with $Y_k = [X_{k+m-1} + u \quad X_{k+m-2} + u \quad \cdots \quad X_k + u]^T$, $a_i = \frac{\partial g}{\partial x_i}\big|_{(c_1, c_2, \cdots, c_m, \xi_0)}$, we can replace the recursion $X_k = g(X_{k-1}, \cdots, X_{k-m}, V_k)$ by the following equivalent recursion:

$$Y_{k+1} = \begin{pmatrix} a_1 & a_2 & \cdots & a_{m-1} & a_m \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{pmatrix} Y_k + \begin{pmatrix} \frac{\partial g}{\partial v}\big|_{(c_1, c_2, \cdots, c_m, \xi_0)} V_k \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}. \tag{80}$$

Since V_k is independent of $\{X_{k+m-i}\}_{i=2}^{m+1}$ or Y_k, (80) models a new 1-order Markov chain $\{Y_k\}_{k=1}^{\infty}$. Then, by using the the same arguments to obtain (73), we can derive bounds on generalization error for this model.

For a general m-order homogeneous Markov chain, it holds that

$$P_{X_k | X_{k-1} = x_1, X_{k-2} = x_2, \cdots, X_{k-m} = x_m} \sim T_{(x_1, x_2, \cdots, x_m)} \tag{81}$$

for all $(x_1, x_2, \cdots, x_m) \in \mathcal{S}^m$, where $T_{(x_1, x_2, \cdots, x_m)}$ is a random variable which depends only on $x_1, x_2, \cdots, x_m$ and does not depend on k. Hence, we can represent the

$$X_k = \tilde{g}(X_{k-1}, X_{k-2}, \cdots, X_{k-m}, T_{(X_{k-1}, X_{k-2}, \cdots, X_{k-m})}), \tag{82}$$

where $T_{(X_{k-1}, X_{k-2}, \cdots, X_{k-m})} = f(\varepsilon_k, V_k, V_{k-1}, \cdots, V_{k-q}, X_{k-1}, X_{k-2}, \cdots, X_{k-m})$. Here, ε_k represents new noise at time k which is independent of the past. Hence, in a general m-order homogeneous Markov chain, we can represent the m-order homogeneous Markov chain by the following recursion:

$$X_k = g\big(X_{k-1}, X_{k-2}, \cdots, X_{k-m}, \varepsilon_k, V_k, V_{k-1}, \cdots, V_{k-q}\big), \tag{83}$$

where ε_k represents new noise at time k and $q \in \mathbb{Z}_+$. By using Taylor expansion to the first order, we can approximate the Markov chain in (83) by an Autoregressive Moving-Average Model (ARMA(m, q)) model as following:

$$X_k = c + \varepsilon_k + \sum_{i=1}^{m} a_i X_{k-i} + \sum_{i=1}^{q} \theta_i \varepsilon_{k-i}, \tag{84}$$

where c and a_i's are constants, and $\{\varepsilon_k\}_{k=1}^{\infty}$ are i. i. d. Gaussian random variables $\mathcal{N}(0, \sigma^2)$. For this model, let

$$V_k := \sum_{i=1}^{k} \varepsilon_i \tag{85}$$

and

$$Y_k := \begin{pmatrix} X_{k+m-1} + u \\ X_{k+m-2} + u \\ \vdots \\ X_k + u \\ V_{k+m-1} \\ V_{k+m-2} \\ \vdots \\ V_{k+m-q} \\ V_{k+m-q-1} \end{pmatrix}, \tag{86}$$

where

$$u := \frac{c}{1 - \sum_{i=1}^{m} a_i}. \tag{87}$$

Let $V_k := \sum_{i=1}^{k} \varepsilon_i$ for all $k \geq 1$. Observe that

$$V_{k+m} = \varepsilon_{k+m} + V_{k+m-1}. \tag{88}$$

On the other hand, we have

$$\begin{aligned} X_{k+m} &= c + \varepsilon_{k+m} + \sum_{i=1}^{m} a_i X_{k+m-i} + \sum_{i=1}^{q} \theta_i \varepsilon_{k+m-i} && (89) \\ &= c + \varepsilon_{k+m} + \sum_{i=1}^{m} a_i X_{k+m-i} + \sum_{i=1}^{q} \theta_i \big(V_{k+m-i} - V_{k+m-1-i}\big) && (90) \\ &= c + \varepsilon_{k+m} + \sum_{i=1}^{m} a_i X_{k+m-i} + \theta_1 V_{k+m-1} + \sum_{i=1}^{q-1} \big(\theta_{i+1} - \theta_i\big) V_{k+m-1-i} - \theta_q V_{k+m-q-1}. && (91) \end{aligned}$$

Then, we have

$$Y_{k+1} = \mathbf{G} Y_k + \begin{pmatrix} \varepsilon_{k+m} \\ 0 \\ \vdots \\ 0 \\ \varepsilon_{k+m} \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \tag{92}$$

where

$$\mathbf{G} := \begin{pmatrix} \mathbf{G}_{11} & \mathbf{G}_{12} \\ \mathbf{G}_{21} & \mathbf{G}_{22} \end{pmatrix}. \tag{93}$$

Here,

$$\mathbf{G}_{11} := \begin{pmatrix} a_1 & a_2 & \cdots & a_{m-1} & a_m \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{pmatrix}_{m \times q+1}, \tag{94}$$

$$\mathbf{G}_{12} := \begin{pmatrix} \theta_1 & \theta_2 - \theta_1 & \cdots & \theta_q - \theta_{q-1} & -\theta_q \\ 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 0 \end{pmatrix}_{m \times q+1}, \tag{95}$$

$$\mathbf{G}_{21} := \begin{pmatrix} 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 0 \end{pmatrix}_{q+1 \times m}, \tag{96}$$

and

$$\mathbf{G}_{22} := \begin{pmatrix} 1 & 0 & \cdots & 0 & 0 \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{pmatrix}_{q+1 \times q+1}. \tag{97}$$

Since ε_{k+m} is independent of Y_k, (92) models a 1-order Markov chain. Hence, we can use the above arguments to derive new generalization error bounds for the m-order homogeneous Markov chain where ARMA model is a special case.

2.2 Mixture of m Services

In this section, we consider the case that $Y_k = \sum_{l=1}^m \alpha_l X_k^{(l)}$ for all $k = 1, 2, \cdots$, where $\{X_k^{(l)}\}_{k=1}^\infty$ are independent Markov chains on $\mathcal{S}$ with stationary distribution for all $l \in [m]$. This setting usually happens in practice, for example, video is a mixture of voice, image, and text, where each service can be modelled as a high-order Markov chain and the order of the Markov chain depends on the type of service.

Let

$$Z_k := \begin{pmatrix} \alpha_1 X_k^{(1)} + \alpha_2 X_k^{(2)} + \cdots \alpha_m X_k^{(m)} \\ \alpha_2 X_k^{(2)} + \cdots \alpha_m X_k^{(m)} \\ \vdots \\ \alpha_m X_k^{(m)} \end{pmatrix} = \begin{pmatrix} Y_k \\ \alpha_2 X_k^{(2)} + \cdots \alpha_m X_k^{(m)} \\ \vdots \\ \alpha_m X_k^{(m)} \end{pmatrix}. \tag{98}$$

Then, it holds that

$$Z_k = \mathbf{G} X_k \tag{99}$$

where

$$\mathbf{G} := \begin{pmatrix} \alpha_1 & \alpha_2 & \cdots & \alpha_{m-1} & \alpha_m \\ 0 & \alpha_2 & \cdots & \alpha_{m-1} & \alpha_m \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & \alpha_m \end{pmatrix}, \tag{100}$$

and

$$X_k := \begin{pmatrix} X_k^{(1)} \\ X_k^{(2)} \\ \vdots \\ X_k^{(m)} \end{pmatrix}. \tag{101}$$

It is obvious that $\mathbf{G}$ is non-singular since $\det(\mathbf{G}) = \prod_{l=1}^m \alpha_l \neq 0$. Therefore, for fixed pair $(x, y) \in \mathcal{S}^m \times \mathcal{S}^m$, we have

$$\mathbb{P}\big(Z_{k+1} = y \big| Z_k = x\big) = \mathbb{P}(X_{k+1} = \mathbf{G}^{-1} y \big| X_k = \mathbf{G}^{-1} x). \tag{102}$$

Now, assume that $\mathbf{G}^{-1}x = \left(\beta_x^{(1)}, \beta_x^{(2)}, \cdots \beta_x^{(m)})\right)$ and $\mathbf{G}^{-1}y = \left(\beta_y^{(1)}, \beta_y^{(2)}, \cdots \beta_y^{(m)})\right)$. Then, from (101) and (102), we have

$$\mathbb{P}\big(Z_{k+1} = y \big| Z_k = x\big) = \mathbb{P}\bigg(\bigcap_{l=1}^m \{X_{k+1}^{(l)} = \beta_y^{(l)}\}\bigg| \bigcap_{l=1}^m \{X_k^{(l)} = \beta_x^{(l)}\}\bigg) \tag{103}$$

$$= \prod_{l=1}^m \mathbb{P}\big(X_{k+1}^{(l)} = \beta_y^{(l)} \big| X_k^{(l)} = \beta_x^{(l)}\big) \tag{104}$$

$$= \prod_{l=1}^m Q_l(\beta_x^{(l)}, \beta_y^{(l)}), \tag{105}$$

where Q_l is the transition probability of the Markov chain l. It follows that $\{Z_k\}_{k=1}^\infty$ is a 1-order Markov chain. It is easy to see that $\{Z_k\}_{k=1}^\infty$ has stationary distribution if all the Markov chains $\{X_k^{(l)}\}_{l=1}^m$ have stationary distributions.

Now, as Subsection 2.1, to derive generalization error bounds for the Markov chain $\{X_n\}_{n=1}^\infty$, we can use the following arguments. For all $f \in \mathcal{F}$ and by setting $\tilde{f}(z_1, z_2, \cdots, z_m) := f(z_1)$ where $\tilde{f} : \mathcal{S}^m \to \mathbb{R}$, we obtain $\tilde{f}(GX_k) = f(Y_k)$ and

$$\frac{1}{n}\sum_{i=1}^n \mathbf{1}\{f(Y_i) \le 0\} = \frac{1}{n}\sum_{i=1}^n \mathbf{1}\{\tilde{f}(\mathbf{G}X_i) \le 0\}. \tag{106}$$

Hence, by applying all the results for 1-order Markov chain $\{Z_n\}_{n=1}^\infty$ where $Z_n = \mathbf{G}X_n$, we obtain corresponding upper bounds for the sequence of m-order Markov chain $\{X_n\}_{n=1}^\infty$.

A Proof of Lemma 1

Before going to prove Lemma 1, we observe the following interesting fact.

Lemma 18. *Let $\{X_n\}_{n=1}^\infty$ be an arbitrary process on a Polish space $\mathcal{S}$, and let $\{Y_n\}_{n=1}^\infty$ be a independent copy (replica) of $\{X_n\}_{n=1}^\infty$. Denote by $\mathbf{X} = (X_1, X_2, \cdots, X_n), \mathbf{Y} = (Y_1, Y_2, \cdots, Y_n)$, and $\mathcal{F}$ a class of uniformly bounded functions from $\mathcal{S} \to \mathbb{R}$. Let $\boldsymbol{\epsilon} := (\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n)$ be a vector of i.i.d. Rademacher's random variables. Then, the following holds:*

$$\mathbb{E}_{\boldsymbol{\epsilon}}\Bigg[\mathbb{E}_{\mathbf{X},\mathbf{Y}}\Bigg[\sup_{f\in\mathcal{F}}\Bigg|\sum_{i=1}^n \varepsilon_i(f(X_i) - f(Y_i))\Bigg|\Bigg]\Bigg] = \mathbb{E}_{\mathbf{X},\mathbf{Y}}\Bigg[\sup_{f\in\mathcal{F}}\Bigg|\sum_{i=1}^n (f(X_i) - f(Y_i))\Bigg|\Bigg]. \tag{107}$$

Remark 19. *Our lemma generalizes a similar fact for i.i.d. processes. In the case that $\{X_n\}_{n=1}^\infty$ is an i.i.d. random process,* (107) *holds with equality since $P_{X^n,Y^n}(x_1, x_2, \cdots, x_n, y_1, y_2, \cdots, y_n)$ is invariant under permutation. However, for the Markov case, this fact does not hold in general. Hence, in the following, we provide a new proof for* (107)*, which works for any process $\{X_n\}_{n=1}^\infty$ by making use of the properties of Rademacher's process.*

Proof. Let $g(x, y) := f(x) - f(y)$ and $\mathcal{G} := \{g : \mathcal{S} \times \mathcal{S} \to \mathbb{R} : g(x, y) := f(x) - f(y) \text{ for some } f \in \mathcal{F}\}$. Then, it holds that

$$\mathbb{E}_{\boldsymbol{\epsilon}}\Bigg[\mathbb{E}_{\mathbf{X},\mathbf{Y}}\Bigg[\sup_{f\in\mathcal{F}}\Bigg|\sum_{i=1}^n \varepsilon_i(f(X_i) - f(Y_i))\Bigg|\Bigg]\Bigg] = \mathbb{E}_{\boldsymbol{\epsilon}}\Bigg[\mathbb{E}_{\mathbf{X},\mathbf{Y}}\Bigg[\sup_{g\in\mathcal{G}}\Bigg|\sum_{i=1}^n \varepsilon_i g(X_i, Y_i)\Bigg|\Bigg]\Bigg]. \tag{108}$$

Observe that $g(X_i, Y_i) = -g(Y_i, X_i)$ for all $i \in [n]$.

For all $j \in [n]$, denote by

$$\mathcal{N}_j := [n] \setminus \{j\}, \tag{109}$$

and

$$\varepsilon_{\mathcal{N}_j} := \{\varepsilon_i : i \in \mathcal{N}_j\}. \tag{110}$$

Now, for each $j \in [n]$, observe that ε_j is independent of $X_1^n, Y_1^n, \varepsilon_{\mathcal{N}_j}$. Hence, we have

$$\begin{aligned}
&\mathbb{E}_{\varepsilon_1^n, X_1^n, Y_1^n}\left[\sup_{g\in\mathcal{G}}\left|\sum_{i=1}^n \varepsilon_i g(X_i,Y_i)\right|\right] \\
&\quad= \frac{1}{2}\mathbb{E}_{\varepsilon_{\mathcal{N}_j}, X_1^n, Y_1^n}\left[\sup_{g\in\mathcal{G}}\left|\sum_{i\in\mathcal{N}_j} \varepsilon_i g(X_i,Y_i)+g(X_j,Y_j)\right|\right] \\
&\qquad+ \frac{1}{2}\mathbb{E}_{\varepsilon_{\mathcal{N}_j}, X_1^n, Y_1^n}\left[\sup_{g\in\mathcal{G}}\left|\sum_{i\in\mathcal{N}_j} \varepsilon_i g(X_i,Y_i)-g(X_j,Y_j)\right|\right] \qquad (111)\\
&\quad= \frac{1}{2}\mathbb{E}_{\varepsilon_{\mathcal{N}_j}, X_1^n, Y_1^n}\left[\sup_{g\in\mathcal{G}}\left|\sum_{i\in\mathcal{N}_j} \varepsilon_i g(X_i,Y_i)+g(X_j,Y_j)\right|\right] \\
&\qquad+ \frac{1}{2}\mathbb{E}_{\varepsilon_{\mathcal{N}_j}, X_1^n, Y_1^n}\left[\sup_{g\in\mathcal{G}}\left|\sum_{i\in\mathcal{N}_j} \varepsilon_i g(X_i,Y_i)+g(Y_j,X_j)\right|\right] \qquad (112)
\end{aligned}$$

where (112) follows from $g(X_j,Y_j) = -g(Y_j,X_j)$.

Now, by setting $\tilde{\epsilon}_i := -\varepsilon_i$ for all $i \in n$. Then, we obtain

$$\begin{aligned}
&\mathbb{E}_{\varepsilon_{\mathcal{N}_j}, X_1^n, Y_1^n}\left[\sup_{g\in\mathcal{G}}\left|\sum_{i\in\mathcal{N}_j} \varepsilon_i g(X_i,Y_i)+g(Y_j,X_j)\right|\right] \\
&\qquad= \mathbb{E}_{\tilde{\varepsilon}_{\mathcal{N}_j}, X_1^n, Y_1^n}\left[\sup_{g\in\mathcal{G}}\left|\sum_{i\in\mathcal{N}_j} \tilde{\epsilon}_i g(X_i,Y_i)+g(Y_j,X_j)\right|\right] \qquad (113)\\
&\qquad= \mathbb{E}_{\varepsilon_{\mathcal{N}_j}, X_1^n, Y_1^n}\left[\sup_{g\in\mathcal{G}}\left|\sum_{i\in\mathcal{N}_j} \varepsilon_i g(Y_i,X_i)+g(Y_j,X_j)\right|\right] \qquad (114)\\
&\qquad= \mathbb{E}_{\varepsilon_{\mathcal{N}_j}, X_1^n, Y_1^n}\left[\sup_{g\in\mathcal{G}}\left|\sum_{i\in\mathcal{N}_j} \varepsilon_i g(X_i,Y_i)+g(X_j,Y_j)\right|\right], \qquad (115)
\end{aligned}$$

where (113) follows from $(\tilde{\epsilon}_i : i \in \mathcal{N}_j)$ has the same distribution as $(\varepsilon_i : i \in \mathcal{N}_j)$, (114) follows from $g(X_i,Y_i) = -g(Y_i,X_i)$ and $\tilde{\epsilon}_i = -\varepsilon_i$, and (115) follows from $g(X_i,Y_i) = -g(Y_i,X_i)$ for all $i \in [n]$.

From (112) and (115), we have

$$\mathbb{E}_{\varepsilon_1^n, X_1^n, Y_1^n}\left[\sup_{g\in\mathcal{G}}\left|\sum_{i=1}^n \varepsilon_i g(X_i,Y_i)\right|\right] = \mathbb{E}_{\varepsilon_{\mathcal{N}_j}, X_1^n, Y_1^n}\left[\sup_{g\in\mathcal{G}}\left|\sum_{i\in\mathcal{N}_j} \varepsilon_i g(X_i,Y_i)+g(X_j,Y_j)\right|\right] \quad \forall j \in [n]. \qquad (116)$$

It follows that

$$\begin{aligned}
&\mathbb{E}_{\varepsilon_1^n, X_1^n, Y_1^n}\left[\sup_{g\in\mathcal{G}}\left|\sum_{i=1}^n \varepsilon_i g(X_i,Y_i)\right|\right] \\
&\quad= \mathbb{E}_{\varepsilon_1^{n-1}, X_1^n, Y_1^n}\left[\sup_{g\in\mathcal{G}}\left|\sum_{i=1}^{n-1} \varepsilon_i g(X_i,Y_i)+g(X_n,Y_n)\right|\right] \qquad (117)\\
&\quad= \frac{1}{2}\mathbb{E}_{\varepsilon_1^{n-2}, X_1^n, Y_1^n}\left[\sup_{g\in\mathcal{G}}\left|\sum_{i=1}^{n-2} \varepsilon_i g(X_i,Y_i)+g(X_{n-1},Y_{n-1})+g(X_n,Y_n)\right|\right] \\
&\qquad+ \frac{1}{2}\mathbb{E}_{\varepsilon_1^{n-2}, X_1^n, Y_1^n}\left[\sup_{g\in\mathcal{G}}\left|\sum_{i=1}^{n-2} \varepsilon_i g(X_i,Y_i)-g(X_{n-1},Y_{n-1})+g(X_n,Y_n)\right|\right] \qquad (118)\\
&\quad= \frac{1}{2}\mathbb{E}_{\varepsilon_1^{n-2}, X_1^n, Y_1^n}\left[\sup_{g\in\mathcal{G}}\left|\sum_{i=1}^{n-2} \varepsilon_i g(X_i,Y_i)+g(X_{n-1},Y_{n-1})+g(X_n,Y_n)\right|\right]
\end{aligned}$$

$$+\frac{1}{2}\mathbb{E}_{\varepsilon_1^{n-2},X_1^n,Y_1^n}\left[\sup_{g\in\mathcal{G}}\left|\sum_{i=1}^{n-2}\varepsilon_i g(X_i,Y_i)+g(Y_{n-1},X_{n-1})+g(X_n,Y_n)\right|\right], \tag{119}$$

where (117) follows from setting $j = n$ in (116), and (119) follows from $g(Y_{n-1},X_{n-1}) = -g(X_{n-1},Y_{n-1})$.

Now, for any fixed tuple $(x_1^{n-1}, y_1^{n-1}, \varepsilon_1^{n-1}) \in \mathcal{S}^{n-1}\times\mathcal{S}^{n-1}\times\{-1,1\}^{n-1}$, observe that

$$\begin{aligned}
&P_{X_n,Y_n|X_1^{n-1},Y_1^{n-1},\varepsilon_1^{n-2}}(x_n,y_n|x_1^{n-1},y_1^{n-1},\varepsilon_1^{n-2})\\
&\quad= P_{X_n|X_1^{n-1}}(x_n|x_1^{n-1})P_{Y_n|Y_1^{n-1}}(y_n|y_1^{n-1}) \qquad (120)\\
&\quad= P_{Y_n|Y_1^{n-1}}(x_n|x_1^{n-1})P_{X_n|X_1^{n-1}}(y_n|y_1^{n-1}) \qquad (121)\\
&\quad= P_{Y_n,X_n|Y_1^{n-1},X_1^{n-1},\varepsilon_1^{n-2}}(x_n,y_n|x_1^{n-1},y_1^{n-1},\varepsilon_1^{n-2}). \qquad (122)
\end{aligned}$$

On the other hand, we also have

$$\begin{aligned}
&P_{X_1^{n-1},Y_1^{n-1},\varepsilon_1^{n-2}}(x_1^{n-1},y_1^{n-1},\varepsilon_1^{n-2})\\
&\quad= P_{\varepsilon_1^{n-2}}(\varepsilon_1^{n-2})P_{X_1^{n-1}}(x_1^{n-1})P_{Y_1^{n-1}}(y_1^{n-1}) \qquad (123)\\
&\quad= P_{\varepsilon_1^{n-2}}(\varepsilon_1^{n-2})P_{Y_1^{n-1}}(x_1^{n-1})P_{X_1^{n-1}}(y_1^{n-1}) \qquad (124)\\
&\quad= P_{Y_1^{n-1},X_1^{n-1},\varepsilon_1^{n-2}}(x_1^{n-1},y_1^{n-1},\varepsilon_1^{n-2}). \qquad (125)
\end{aligned}$$

Hence, from (122) and (125), we obtain

$$\begin{aligned}
&P_{X_1^n,Y_1^n,\varepsilon_1^{n-2}}\big(x_1^n,y_1^n,\varepsilon_1^{n-2}\big)\\
&\quad= P_{X_1^{n-1}Y_1^{n-1}\varepsilon_1^{n-2}}(x_1^{n-1},y_1^{n-1},\varepsilon_1^{n-2})P_{X_nY_n|X_1^{n-1}Y_1^{n-1}\varepsilon_1^{n-2}}\big(x_n,y_n|x_1^{n-1},y_1^{n-1},\varepsilon_1^{n-2}\big) \qquad (126)\\
&\quad= P_{Y_1^{n-1},X_1^{n-1},\varepsilon_1^{n-2}}(x_1^{n-1},y_1^{n-1},\varepsilon_1^{n-2})P_{Y_n,X_n|Y_1^{n-1},X_1^{n-1},\varepsilon_1^{n-2}}\big(x_n,y_n|x_1^{n-1},y_1^{n-1},\varepsilon_1^{n-2}\big) \qquad (127)\\
&\quad= P_{Y_1^n,X_1^n,\varepsilon_1^{n-2}}\big(x_1^n,y_1^n,\varepsilon_1^{n-2}\big). \qquad (128)
\end{aligned}$$

Now, from (128), we also have

$$P_{X_1^n,Y_1^n}(x_1^n,y_1^n) = P_{Y_1^nX_1^n}(x_1^n,y_1^n). \tag{129}$$

It follows from (129) that

$$\begin{aligned}
P_{X_{n-1},Y_{n-1}}(x_{n-1},y_{n-1}) &= \sum_{x_1^{n-2},y_1^{n-2},x_n,y_n} P_{X_1^n,Y_1^n}(x_1^n,y_1^n) \qquad (130)\\
&= \sum_{x_1^{n-2},y_1^{n-2},x_n,y_n} P_{Y_1^nX_1^n}(x_1^n,y_1^n) \qquad (131)\\
&= P_{Y_{n-1}X_{n-1}}(x_{n-1},y_{n-1}). \qquad (132)
\end{aligned}$$

Hence, from (128) and (132), we have

$$\begin{aligned}
&P_{X_n,Y_n,X_1^{n-2},Y_1^{n-2},\varepsilon_1^{n-2}|X_{n-1},Y_{n-1}}\big(x_n,y_n,x_1^{n-2},y_1^{n-2},\varepsilon_1^{n-2}|x_{n-1},y_{n-1}\big) \qquad (133)\\
&\quad= \frac{P_{X_1^n,Y_1^n,\varepsilon_1^{n-2}}(x_1^n,y_1^n,\varepsilon_1^{n-2})}{P_{X_{n-1}Y_{n-1}}(x_{n-1},y_{n-1})} \qquad (134)\\
&\quad= \frac{P_{Y_1^n,X_1^n,\varepsilon_1^{n-2}}\big(x_1^n,y_1^n,\varepsilon_1^{n-2}\big)}{P_{Y_{n-1}X_{n-1}}(x_{n-1},y_{n-1})} \qquad (135)\\
&\quad= P_{X_n,Y_n,X_1^{n-2},Y_1^{n-2},\varepsilon_1^{n-2}|X_{n-1},Y_{n-1}}\big(y_n,x_n,y_1^{n-2},x_1^{n-2},\varepsilon_1^{n-2}|y_{n-1},x_{n-1}\big). \qquad (136)
\end{aligned}$$

From (136), for each fixed $(x_{n-1}, y_{n-1}) \in \mathcal{S} \times \mathcal{S}$, we have

$$\begin{aligned}
&\mathbb{E}_{\varepsilon_1^{n-2},X_1^{n-2},Y_1^{n-2},X_n,Y_n}\left[\sup_{g\in\mathcal{G}}\left|\sum_{i=1}^{n-2}\varepsilon_i g(X_i,Y_i) + g(Y_{n-1},X_{n-1})\right.\right. \\
&\qquad\qquad \left.\left. + g(X_n,Y_n)\right| \middle| X_{n-1}=x_{n-1}, Y_{n-1}=y_{n-1}\right] \\
&\quad = \mathbb{E}_{\varepsilon_1^{n-2},X_1^{n-2},Y_1^{n-2},X_n,Y_n}\left[\sup_{g\in\mathcal{G}}\left|\sum_{i=1}^{n-2}\varepsilon_i g(X_i,Y_i) + g(y_{n-1},x_{n-1})\right.\right. \\
&\qquad\qquad \left.\left. + g(X_n,Y_n)\right| \middle| X_{n-1}=x_{n-1}, Y_{n-1}=y_{n-1}\right] \\
&\quad = \mathbb{E}_{\varepsilon_1^{n-2},X_1^{n-2},Y_1^{n-2},X_n,Y_n}\left[\sup_{g\in\mathcal{G}}\left|\sum_{i=1}^{n-2}\varepsilon_i g(Y_i,X_i) + g(y_{n-1},x_{n-1})\right.\right. \\
&\qquad\qquad \left.\left. + g(Y_n,X_n)\right| \middle| Y_{n-1}=x_{n-1}, X_{n-1}=y_{n-1}\right] \qquad (137) \\
&\quad = \mathbb{E}_{\varepsilon_1^{n-2},X_1^{n-2},Y_1^{n-2},X_n,Y_n}\left[\sup_{g\in\mathcal{G}}\left|\sum_{i=1}^{n-2}\varepsilon_i g(X_i,Y_i) + g(x_{n-1},y_{n-1})\right.\right. \\
&\qquad\qquad \left.\left. + g(X_n,Y_n)\right| \middle| Y_{n-1}=x_{n-1}, X_{n-1}=y_{n-1}\right] \qquad (138)
\end{aligned}$$

where (137) follows from (136), and (138) follows from the fact that $g(x,y) = -g(y,x)$ for all $x, y \in \mathcal{S}\times\mathcal{S}$.

From (138) and (132), we obtain

$$\begin{aligned}
&\mathbb{E}_{\varepsilon_1^{n-2},X_1^n,Y_1^n}\left[\sup_{g\in\mathcal{G}}\left|\sum_{i=1}^{n-2}\varepsilon_i g(X_i,Y_i) + g(Y_{n-1},X_{n-1}) + g(X_n,Y_n)\right|\right] \\
&\quad = \mathbb{E}_{\varepsilon_1^{n-2},X_1^n,Y_1^n}\left[\sup_{g\in\mathcal{G}}\left|\sum_{i=1}^{n-2}\varepsilon_i g(X_i,Y_i) + g(X_{n-1},Y_{n-1}) + g(X_n,Y_n)\right|\right]. \qquad (139)
\end{aligned}$$

From (119) and (139), we obtain

$$\begin{aligned}
&\mathbb{E}_{\varepsilon_1^{n},X_1^n,Y_1^n}\left[\sup_{g\in\mathcal{G}}\left|\sum_{i=1}^{n}\varepsilon_i g(X_i,Y_i)\right|\right] \\
&\quad = \mathbb{E}_{\varepsilon_1^{n-2},X_1^n,Y_1^n}\left[\sup_{g\in\mathcal{G}}\left|\sum_{i=1}^{n-2}\varepsilon_i g(X_i,Y_i) + g(X_{n-1},Y_{n-1}) + g(X_n,Y_n)\right|\right]. \qquad (140)
\end{aligned}$$

By using induction, we finally obtain

$$\mathbb{E}_{\varepsilon_1^{n},X_1^n,Y_1^n}\left[\sup_{g\in\mathcal{G}}\left|\sum_{i=1}^{n}\varepsilon_i g(X_i,Y_i)\right|\right] = \mathbb{E}_{X_1^n,Y_1^n}\left[\sup_{g\in\mathcal{G}}\left|\sum_{i=1}^{n} g(X_i,Y_i)\right|\right], \qquad (141)$$

or equation (107) holds. □

Next, recall the following result which was developed base on the spectral method [6]:

Lemma 20. *[7, Theorems 3.41] Let $X_1, X_2, \cdots, X_n$ be a stationary Markov chain on some Polish space with L_2 spectral gap λ defined in Section 2.1 in the main document and the initial distribution $\nu \in \mathcal{M}_2$. Let $f \in \mathcal{F}$ and define*

$$S_{n,n_0}(f) = \frac{1}{n}\sum_{j=1}^{n} f(X_{j+n_0}) \qquad (142)$$

for all $n_0 \geq 0$. Then, it holds that

$$\mathbb{E}\left[\left|S_{n,n_0}(f) - \mathbb{E}_\pi[f(X)]\right|^2\right] \leq \frac{2M}{n(1-\lambda)} + \frac{64M^2}{n^2(1-\lambda)^2}\lambda^{n_0}\left\|\frac{dv}{d\pi} - 1\right\|_2. \qquad (143)$$

In addition, we recall the following McDiarmid's inequality for Markov chain:

Lemma 21. *[10, Cor. 2.10] Let* $\mathbf{X} := (X_1, X_2, \cdots, X_n)$ *be a homogeneous Markov chain, taking values in a Polish state space* $\Lambda = \underbrace{\mathcal{S} \times \mathcal{S} \times \cdots \times \mathcal{S}}_{n \text{ times}}$*, with mixing time* $t_{\text{mix}}(\varepsilon)$ *(for* $0 \leq \varepsilon \leq 1$*). Recall the definition of* $\tau_{\min}$ *in Eq. (3) in the main document. Suppose that* $f : \Lambda \to \mathbb{R}$ *satisfies*

$$f(\mathbf{x}) - f(\mathbf{y}) \leq \sum_{i=1}^n c_i \mathbf{1}\{x_i \neq y_i\} \tag{144}$$

for every $\mathbf{x}, \mathbf{y} \in \Lambda$*, for some* $c \in \mathbb{R}_+^n$*. Then, for any* $t \geq 0$*,*

$$\mathbb{P}_{\mathbf{X}}\big[\big|f(\mathbf{X}) - \mathbb{E}[f(\mathbf{X})]\big| \geq t\big] \leq 2\exp\bigg(-\frac{2t^2}{\|c\|_2^2 \tau_{\min}}\bigg). \tag{145}$$

Now, we return to the proof of Lemma 1.

Proof of Lemma 1. For each $f \in \mathcal{F}$, observe that

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^n f(X_i) - \int_{\mathcal{S}} \pi(x) f(x) dx \\ &= \frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}[f(X_i)] + \frac{1}{n}\sum_{i=1}^n \mathbb{E}[f(X_i)] - \int_{\mathcal{S}} \pi(x) f(x) dx. \end{aligned} \tag{146}$$

On the other hand, we have

$$\begin{aligned}&\bigg|\frac{1}{n}\sum_{i=1}^n \mathbb{E}[f(X_i)] - \int_{\mathcal{S}} \pi(x) f(x) dx\bigg| \\ &\qquad = \mathbb{E}\bigg[\bigg|S_{n,0}(f) - \mathbb{E}_\pi[f(X)]\bigg|\bigg] && (147)\\ &\qquad \leq \sqrt{\mathbb{E}\bigg[\bigg|S_{n,0}(f) - \mathbb{E}_\pi[f(X)]\bigg|^2\bigg]} && (148)\\ &\qquad \leq \sqrt{\frac{2M}{n(1-\lambda)} + \frac{64M^2}{n^2(1-\lambda)^2}\bigg\|\frac{dv}{d\pi} - 1\bigg\|_2} && (149)\\ &\qquad = A_n, && (150)\end{aligned}$$

where (150) follows from Lemma 20 with $n_0 = 0$.

By using $|a+b| \leq |a| + |b$, from (146) and (150), we obtain

$$\begin{aligned}&\mathbb{E}\big[\big\|P_n - P\big\|_{\mathcal{F}}\big] \\ &\qquad \leq \mathbb{E}\bigg[\sup_{f\in\mathcal{F}}\bigg|\frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}[f(X_i)]\bigg|\bigg] + A_n.\end{aligned} \tag{151}$$

On the other hand, let $Y_1, Y_2, \cdots, Y_n$ is a replica of $X_1, X_2, \cdots, X_n$. It holds that

$$\begin{aligned}&\mathbb{E}\bigg[\sup_{f\in\mathcal{F}}\bigg|\frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}[f(X_i)]\bigg|\bigg] \\ &\qquad = \mathbb{E}_{\mathbf{X}}\bigg[\sup_{f\in\mathcal{F}}\bigg|\frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}[f(Y_i)]\bigg|\bigg] && (152)\\ &\qquad = \mathbb{E}_{\mathbf{X}}\bigg[\sup_{f\in\mathcal{F}}\bigg|\mathbb{E}_{\mathbf{Y}}\bigg[\frac{1}{n}\sum_{i=1}^n f(X_i) - f(Y_i)\bigg]\bigg|\bigg] && (153)\\ &\qquad \leq \mathbb{E}_{\mathbf{X}}\bigg[\mathbb{E}_{\mathbf{Y}}\bigg[\sup_{f\in\mathcal{F}}\bigg|\frac{1}{n}\sum_{i=1}^n f(X_i) - f(Y_i)\bigg|\bigg]\bigg]. && (154)\end{aligned}$$

Now, by Lemma 18 and the triangle inequality for infinity norm, we have

$$\mathbb{E}\bigg[\sup_{f\in\mathcal{F}}\bigg|\frac{1}{n}\sum_{i=1}^n f(X_i)-f(Y_i)\bigg|\bigg]=\mathbb{E}_\varepsilon\mathbb{E}_{\mathbf{X},\mathbf{Y}}\bigg[\bigg\|\frac{1}{n}\sum_{i=1}^n\varepsilon_i\big(f(X_i)-f(Y_i)\big)\bigg\|_{\mathcal{F}}\bigg] \tag{155}$$

$$\leq \mathbb{E}_\varepsilon\mathbb{E}_{\mathbf{X}}\bigg[\bigg\|\frac{1}{n}\sum_{i=1}^n\varepsilon_i\big(f(X_i)\big)\bigg\|_{\mathcal{F}}\bigg]+\mathbb{E}_\varepsilon\mathbb{E}_{\mathbf{Y}}\bigg[\bigg\|\frac{1}{n}\sum_{i=1}^n\varepsilon_i\big(f(Y_i)\big)\bigg\|_{\mathcal{F}}\bigg] \tag{156}$$

$$=2\mathbb{E}\big[\|P_n^0\|_{\mathcal{F}}\big], \tag{157}$$

where (157) follows from the fact that $\mathbf{Y}$ is a replica of $\mathbf{X}$.

From (151) and (157), we finally obtain

$$\mathbb{E}\big[\big\|P_n-P\big\|_{\mathcal{F}}\big]\leq 2\mathbb{E}\big[\|P_n^0\|_{\mathcal{F}}\big]+A_n. \tag{158}$$

Now, by using the triangle inequality, we have

$$\mathbb{E}[\|P_n^0\|_{\mathcal{F}}]=\mathbb{E}_{\mathbf{X},\boldsymbol{\varepsilon}}\bigg[\sup_{f\in\mathcal{F}}\bigg|\frac{1}{n}\sum_{i=1}^n\varepsilon_i f(X_i)\bigg|\bigg] \tag{159}$$

$$\leq \mathbb{E}\bigg[\sup_{f\in\mathcal{F}}\bigg|\frac{1}{n}\sum_{i=1}^n\varepsilon_i\big(f(X_i)-\mathbb{E}_{\mathbf{Y}}[f(Y_i)]\big)\bigg|+\mathbb{E}\bigg[\sup_{f\in\mathcal{F}}\bigg|\frac{1}{n}\sum_{i=1}^n\varepsilon_i\mathbb{E}_{\mathbf{Y}}[f(Y_i)]\bigg|\bigg] \tag{160}$$

$$=\mathbb{E}_{\mathbf{X},\boldsymbol{\varepsilon}}\bigg[\sup_{f\in\mathcal{F}}\bigg|\mathbb{E}_{\mathbf{Y}}\bigg[\frac{1}{n}\sum_{i=1}^n\varepsilon_i\big(f(X_i)-f(Y_i)\big)\bigg]\bigg|\bigg]+\mathbb{E}\bigg[\sup_{f\in\mathcal{F}}\bigg|\frac{1}{n}\sum_{i=1}^n\varepsilon_i\mathbb{E}_{\mathbf{Y}}[f(Y_i)]\bigg|\bigg] \tag{161}$$

$$\leq \mathbb{E}_{\mathbf{X},\mathbf{Y},\boldsymbol{\varepsilon}}\bigg[\sup_{f\in\mathcal{F}}\bigg|\frac{1}{n}\sum_{i=1}^n\varepsilon_i\big(f(X_i)-f(Y_i)\big)\bigg|\bigg]+\mathbb{E}\bigg[\sup_{f\in\mathcal{F}}\bigg|\frac{1}{n}\sum_{i=1}^n\varepsilon_i\mathbb{E}_{\mathbf{Y}}[f(Y_i)]\bigg|\bigg] \tag{162}$$

$$=\mathbb{E}_{\mathbf{X},\mathbf{Y}}\bigg[\sup_{f\in\mathcal{F}}\bigg|\frac{1}{n}\sum_{i=1}^n\big(f(X_i)-f(Y_i)\big)\bigg|\bigg]+\mathbb{E}\bigg[\sup_{f\in\mathcal{F}}\bigg|\frac{1}{n}\sum_{i=1}^n\varepsilon_i\mathbb{E}_{\mathbf{Y}}[f(Y_i)]\bigg|\bigg], \tag{163}$$

where (163) follows from Lemma 18.

Now, we have

$$\mathbb{E}_{\mathbf{X},\mathbf{Y}}\bigg[\sup_{f\in\mathcal{F}}\bigg|\frac{1}{n}\sum_{i=1}^n\big(f(X_i)-f(Y_i)\big)\bigg|\bigg]$$

$$=\mathbb{E}_{\mathbf{X},\mathbf{Y}}\bigg[\sup_{f\in\mathcal{F}}\bigg|\frac{1}{n}\sum_{i=1}^n\big(f(X_i)-Pf\big)-\big(f(Y_i)-Pf\big)\big)\bigg|\bigg] \tag{164}$$

$$\leq \mathbb{E}_{\mathbf{X},\boldsymbol{\varepsilon}}\bigg[\sup_{f\in\mathcal{F}}\bigg|\frac{1}{n}\sum_{i=1}^n f(X_i)-Pf\bigg|\bigg]+\mathbb{E}_{\mathbf{Y},\boldsymbol{\varepsilon}}\bigg[\sup_{f\in\mathcal{F}}\bigg|\frac{1}{n}\sum_{i=1}^n f(Y_i)-Pf\bigg|\bigg] \tag{165}$$

$$=2\mathbb{E}_{\mathbf{X},\boldsymbol{\varepsilon}}\bigg[\sup_{f\in\mathcal{F}}\bigg|\frac{1}{n}\sum_{i=1}^n f(X_i)-Pf\bigg|\bigg] \tag{166}$$

$$=2\mathbb{E}\big[\|P_n-P\|_{\mathcal{F}}\big]. \tag{167}$$

On the other hand, by using the triangle inequality, we also have

$$\mathbb{E}\bigg[\sup_{f\in\mathcal{F}}\bigg|\frac{1}{n}\sum_{i=1}^n\varepsilon_i\mathbb{E}_{\mathbf{Y}}[f(Y_i)]\bigg|\bigg]$$

$$\leq \mathbb{E}\bigg[\sup_{f\in\mathcal{F}}\bigg|\frac{1}{n}\sum_{i=1}^n\varepsilon_i\big(\mathbb{E}_{\mathbf{Y}}[f(Y_i)]-\mathbb{E}_\pi[f(Y)]\big)\bigg|\bigg]+\mathbb{E}\bigg[\sup_{f\in\mathcal{F}}\bigg|\frac{1}{n}\sum_{i=1}^n\varepsilon_i\mathbb{E}_\pi[f(Y)]\bigg|\bigg]. \tag{168}$$

Now, observe that

$$\mathbb{E}\bigg[\sup_{f\in\mathcal{F}}\bigg|\frac{1}{n}\sum_{i=1}^n \varepsilon_i \mathbb{E}_\pi[f(Y)]\bigg|\bigg] \leq \bigg(\sup_{f\in\mathcal{F}}\big|\mathbb{E}_\pi[f(Y)]\big|\bigg)\mathbb{E}_{\boldsymbol{\varepsilon}}\bigg[\sum_{i=1}^n \frac{1}{n}\sum_{i=1}^n \varepsilon_i\bigg] \tag{169}$$

$$\leq M\mathbb{E}_{\boldsymbol{\varepsilon}}\bigg[\sum_{i=1}^n \frac{1}{n}\sum_{i=1}^n \varepsilon_i\bigg] \tag{170}$$

$$\leq M\sqrt{\mathbb{E}_{\boldsymbol{\varepsilon}}\bigg[\bigg(\frac{1}{n}\sum_{i=1}^n \varepsilon_i\bigg)^2\bigg]} \tag{171}$$

$$= \frac{M}{\sqrt{n}}. \tag{172}$$

In addition, for each fixed $(\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n) \in \{-1,+1\}^n$ and $f \in \mathcal{F}$, we have

$$\bigg|\frac{1}{n}\sum_{i=1}^n \varepsilon_i\big(\mathbb{E}_{\mathbf{Y}}[f(Y_i)] - \mathbb{E}_\pi[f(Y)]\big)\bigg|$$

$$\leq \frac{1}{n}\mathbb{E}_{\mathbf{Y}}\bigg|\sum_{i=1}^n \varepsilon_i\big(f(Y_i) - \mathbb{E}_\pi[f(Y)]\big)\bigg| \tag{173}$$

$$= \frac{1}{n}\mathbb{E}_{\mathbf{Y}}\big[\big|g_\varepsilon(\mathbf{Y})\big|\big], \tag{174}$$

where

$$g_\varepsilon(\mathbf{y}) := \sum_{i=1}^n \varepsilon_i\big(f(y_i) - \mathbb{E}_\pi[f(Y)]\big), \qquad \forall \mathbf{y} \in \mathbb{R}^n. \tag{175}$$

Now, for all $\mathbf{x}, \mathbf{y} \in \mathbf{b}^n$, we have

$$g_\varepsilon(\mathbf{x}) - g_\varepsilon(\mathbf{y}) = \sum_{i=1}^n \varepsilon_i\big(f(x_i) - f(y_i)\big) \tag{176}$$

$$\leq \sum_{i=1}^n |f(x_i) - f(y_i)| \tag{177}$$

$$\leq \sum_{i=1}^n 2M\mathbf{1}\{x_i \neq y_i\}. \tag{178}$$

Hence, by Lemma 21, we have

$$\mathbb{P}_{\mathbf{Y}}\big[\big|g_\varepsilon(\mathbf{Y})\big| \geq M\sqrt{2\tau_{\min} n \log n}\big] \leq \frac{2}{n}. \tag{179}$$

It follows that

$$\mathbb{E}_{\mathbf{Y}}\big[\big|g_\varepsilon(\mathbf{Y})\big|\big] \leq \mathbb{E}_{\mathbf{Y}}\big[\big|g_\varepsilon(\mathbf{Y})\big|\big|\big|g_\varepsilon(\mathbf{Y})\big| < M\sqrt{2\tau_{\min} n \log n}\big]$$
$$+ 2nM\mathbb{P}_{\mathbf{Y}}\big[\big|g_\varepsilon(\mathbf{Y})\big| \geq M\sqrt{2\tau_{\min} n \log n}\big] \tag{180}$$

$$\leq M\sqrt{2\tau_{\min} n \log n} + 2nM\bigg(\frac{2}{n}\bigg) \tag{181}$$

$$= M\sqrt{2\tau_{\min} n \log n} + 4M, \tag{182}$$

where (180) follows from the total law of expectation and $|g_\varepsilon(\mathbf{y})| \leq \sum_{i=1}^n \big|f(y_i) - \mathbb{E}_\pi[f(Y)]\big| \leq 2nM$ for all $\mathbf{y} \in \mathbb{R}^n$.

From (174) and (182), we have

$$\bigg|\frac{1}{n}\sum_{i=1}^n \varepsilon_i\big(\mathbb{E}_{\mathbf{Y}}[f(Y_i)] - \mathbb{E}_\pi[f(Y)]\big)\bigg| \leq \frac{1}{n}\bigg(M\sqrt{2\tau_{\min} n \log n} + 4M\bigg) \tag{183}$$

for all $f \in \mathcal{F}$, $(\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n) \in \{-1, +1\}^n$.

From (183), we obtain

$$\mathbb{E}_{\boldsymbol{\varepsilon}}\bigg[\sup_{f \in \mathcal{F}}\bigg|\frac{1}{n}\sum_{i=1}^n \varepsilon_i\big(\mathbb{E}_{\mathbf{Y}}[f(Y_i)] - \mathbb{E}_\pi[f(Y)]\big)\bigg|\bigg] \le \frac{1}{n}\bigg(M\sqrt{2\tau_{\min} n \log n} + 4M\bigg). \tag{184}$$

From (168), (172), and (184), we have

$$\begin{aligned}&\mathbb{E}\bigg[\sup_{f \in \mathcal{F}}\bigg|\frac{1}{n}\sum_{i=1}^n \varepsilon_i \mathbb{E}_{\mathbf{Y}}[f(Y_i)]\bigg|\bigg] \\ &\qquad \le \frac{1}{n}\bigg(M\sqrt{2\tau_{\min} n \log n} + 4M\bigg) + \frac{M}{\sqrt{n}}. \end{aligned}\tag{185}$$

From (163), (167), and (185), we obtain

$$\mathbb{E}[\|P_n^0\|_{\mathcal{F}}] \le 2\mathbb{E}\big[\|P_n - P\|_{\mathcal{F}}\big] + \frac{1}{n}\bigg(M\sqrt{2\tau_{\min} n \log n} + 4M\bigg) + \frac{M}{\sqrt{n}}, \tag{186}$$

and we finally have

$$\mathbb{E}\big[\|P_n - P\|_{\mathcal{F}}\big] \ge \frac{1}{2}\mathbb{E}[\|P_n^0\|_{\mathcal{F}}] - \tilde{A}_n. \tag{187}$$

□

B Proof of Theorem 2

The proof of Theorem 2 is based on [2]. First, we prove (6). Without loss of generality, we can assume that each $\varphi \in \Phi$ takes its values in $[0, 1]$ (otherwise, it can be redefined as $\varphi \wedge 1$). Then, it is clear that $\varphi(x) = 1$ for $x \le 0$. Hence, for each fixed $\varphi \in \Phi$ and $f \in \mathcal{F}$, we obtain

$$\begin{aligned} P\{f \le 0\} &\le P\varphi(f) \qquad (188) \\ &\le P_n\varphi(f) + \|P_n - P\|_{\mathcal{G}_\varphi}, \qquad (189)\end{aligned}$$

where

$$\mathcal{G}_\varphi := \big\{\varphi \cdot f : f \in \mathcal{F}\big\}. \tag{190}$$

Now, let $g(\mathbf{x}) = \sup_{f \in \mathcal{G}_\varphi}\big|\frac{1}{n}\sum_{i=1}^n f(x_i) - Pf\big|$. Then, for all $\mathbf{x} \neq \mathbf{y}$, we have

$$\begin{aligned}\big|g(\mathbf{x}) - g(\mathbf{y})\big| &= \bigg|\sup_{f \in \mathcal{G}_\varphi}\bigg|\frac{1}{n}\sum_{i=1}^n f(x_i) - Pf\bigg| - \sup_{f \in \mathcal{G}_\varphi}\bigg|\frac{1}{n}\sum_{i=1}^n f(y_i) - Pf\bigg|\bigg| \qquad (191)\\ &\le \sup_{f \in \mathcal{G}_\varphi}\bigg|\bigg|\frac{1}{n}\sum_{i=1}^n f(x_i) - Pf\bigg| - \bigg|\frac{1}{n}\sum_{i=1}^n f(y_i) - Pf\bigg|\bigg| \qquad (192)\\ &\le \sup_{f \in \mathcal{G}_\varphi}\bigg|\bigg(\frac{1}{n}\sum_{i=1}^n f(x_i) - Pf\bigg) - \bigg(\frac{1}{n}\sum_{i=1}^n f(y_i) - Pf\bigg)\bigg| \qquad (193)\\ &\le \sup_{f \in \mathcal{G}_\varphi}\bigg|\frac{1}{n}\sum_{i=1}^n f(x_i) - f(y_i)\bigg| \qquad (194)\\ &\le \sup_{f \in \mathcal{G}_\varphi}\frac{1}{n}\sum_{i=1}^n \big|f(x_i) - f(y_i)\big| \qquad (195)\\ &\le \frac{1}{n}\sum_{i=1}^n \mathbf{1}\{x_i \neq y_i\}. \qquad (196)\end{aligned}$$

Hence, for $t > 0$, by Lemma 21 with we have

$$\mathbb{P}\bigg\{\|P_n - P\|_{\mathcal{G}_\varphi} \ge \mathbb{E}\big[\|P_n - P\|_{\mathcal{G}_\varphi}\big] + t\sqrt{\frac{\tau_{\min}}{n}}\bigg\} \le 2\exp\big(-2t^2\big). \tag{197}$$

Hence, with probability at least $1-2\exp\big(-2t^2\big)$ for all $f \in \mathcal{F}$

$$P\{f \le 0\} \le P_n\varphi(f) + \mathbb{E}[\|P_n - P\|_{\mathcal{G}_\varphi}] + t\sqrt{\frac{\tau_{\min}}{n}}. \tag{198}$$

Now, by Lemma 1 with $M = 1$ (since $\sup_{f\in\mathcal{G}_\varphi}\|f\|_\infty = 1$), it holds that

$$\mathbb{E}\Big[\big\|P_n - P\big\|_{\mathcal{G}_\varphi}\Big] \le 2\mathbb{E}[\|P_n^0\|_{\mathcal{G}_\varphi}] + A_n\big|_{M=1} \tag{199}$$

$$= 2\mathbb{E}\Bigg[\Bigg\|n^{-1}\sum_{i=1}^n \varepsilon_i\delta_{X_i}\Bigg\|_{\mathcal{G}_\varphi}\Bigg] + B_n. \tag{200}$$

Since $(\varphi-1)/L(\varphi)$ is contractive and $\varphi(0)-1=0$, by using Talagrand's contraction lemma [3, 11], we obtain

$$\mathbb{E}_\varepsilon\Bigg\|n^{-1}\sum_{i=1}^n \varepsilon_i\delta_{X_i}\Bigg\|_{\mathcal{G}_\varphi} \le 2L(\varphi)\mathbb{E}_\varepsilon\Bigg\|n^{-1}\sum_{i=1}^n \varepsilon_i\delta_{X_i}\Bigg\|_{\mathcal{F}} \tag{201}$$

$$= 2L(\varphi)R_n(\mathcal{F}). \tag{202}$$

From (197), (198), (200), and (202), with probability $1-2\exp\big(-2t^2\big)$, we have for all $f \in \mathcal{F}$,

$$P\{f \le 0\} \le P_n\varphi(f) + 4L(\varphi)R_n(\mathcal{F}) + t\sqrt{\frac{\tau_{\min}}{n}} + B_n. \tag{203}$$

Now, we use (203) with $\varphi = \varphi_k$ and t is replaced by $t + \sqrt{\log k}$ to obtain

$$\mathbb{P}\Bigg(\exists f \in \mathcal{F} : P\{f \le 0\} > \inf_{k>0}\Bigg[P_n\varphi_k(f) + 4L(\varphi_k)R_n(\mathcal{F}) + \big(t+\sqrt{\log k}\big)\sqrt{\frac{\tau_{\min}}{n}} + B_n\Bigg]\Bigg)$$

$$\le 2\sum_{k=1}^\infty \exp\Big(-2\big(t+\sqrt{\log k}\big)^2\Big) \tag{204}$$

$$\le 2\sum_{k=1}^\infty k^{-2}\exp\big(-2t^2\big) \tag{205}$$

$$= \frac{\pi^2}{3}\exp\big(-2t^2\big), \tag{206}$$

where (206) follows from

$$\frac{\pi^2}{6} = \sum_{k=1}^\infty k^{-2}. \tag{207}$$

Next, we prove (7). By the equivalence of Rademacher and Gaussian complexity [12], we have

$$\mathbb{E}\Bigg\|n^{-1}\sum_{i=1}^n \varepsilon_i\delta_{X_i}\Bigg\|_{\mathcal{G}_\varphi} \le \sqrt{\frac{\pi}{2}}\mathbb{E}\Bigg\|n^{-1}\sum_{i=1}^n g_i\delta_{X_i}\Bigg\|_{\mathcal{G}_\varphi}. \tag{208}$$

Hence, from (200) and (208), we obtain

$$\mathbb{E}\Big[\big\|P_n - P\big\|_{\mathcal{G}_\varphi}\Big] \le \sqrt{2\pi}\mathbb{E}\Bigg\|n^{-1}\sum_{i=1}^n g_i\delta_{X_i}\Bigg\|_{\mathcal{G}_\varphi} + B_n. \tag{209}$$

Now, define Gaussian processes

$$Z_1(f,\sigma) := \sigma n^{-1/2}\sum_{i=1}^n g_i(\varphi\circ f)(X_i), \tag{210}$$

and

$$Z_2(f,\sigma) := L(\varphi)n^{-1/2}\sum_{i=1}^n g_i f(X_i) + \sigma g, \tag{211}$$

where $\sigma = \pm 1$ and g is standard normal independent of the sequence $\{g_i\}$. Let $\mathbb{E}_g$ be the expectation on the probability space $(\Omega_g, \Sigma_g, \mathbb{P}_g)$ on which the sequence $\{g_i\}$ and g are defined, then by [2, 11], we have

$$\mathbb{E}_g\biggl[\sup\{Z_1(f,\sigma) : f \in \mathcal{F}, \sigma = \pm 1\}\biggr] \le \mathbb{E}_g\biggl[\sup\{Z_2(f,\sigma) : f \in \mathcal{F}, \sigma = \pm 1\}\biggr]. \tag{212}$$

On the other hand, it holds that

$$\mathbb{E}_g\biggl\|n^{-1/2}\sum_{i=1}^n g_i \delta_{X_i}\biggr\|_{\mathcal{G}_\varphi} = \mathbb{E}_g\biggl[n^{-1/2}\sup_{h \in \bar{\mathcal{G}}_\varphi}\sum_{i=1}^n g_i h(X_i)\biggr] \tag{213}$$

$$= \mathbb{E}_g\biggl[\sup\{Z_1(f,\sigma) : f \in \mathcal{F}, \sigma = \pm 1\}\biggr], \tag{214}$$

where $\bar{\mathcal{G}}_\varphi := \{\varphi(f), -\varphi(f) : f \in \mathcal{F}\}$, and similarly

$$L(\varphi)\mathbb{E}_g\biggl\|n^{-1/2}\sum_{i=1}^n g_i \delta_{X_i}\biggr\|_{\mathcal{F}} + \mathbb{E}|g| \ge \mathbb{E}_g\biggl[\sup\{Z_2(f,\sigma) : f \in \mathcal{F}, \sigma = \pm 1\}\biggr]. \tag{215}$$

From (212), (214), and (215), we have

$$\mathbb{E}_g\biggl\|n^{-1}\sum_{i=1}^n g_i \delta_{X_i}\biggr\|_{\mathcal{G}_\varphi} \le L(\varphi)\mathbb{E}_g\biggl\|n^{-1}\sum_{i=1}^n g_i \delta_{X_i}\biggr\|_{\mathcal{F}} + n^{-1/2}\mathbb{E}|g|. \tag{216}$$

By combining (209) and (216), we obtain

$$\mathbb{E}\bigl[\|P_n - P\|_{\mathcal{G}_\varphi}\bigr] \qquad \le \sqrt{2\pi}\biggl(L(\varphi)\mathbb{E}\biggl[\biggl\|n^{-1}\sum_{i=1}^n g_i \delta_{X_i}\biggr\|_{\mathcal{F}}\biggr] + n^{-1/2}\mathbb{E}|g|\biggr) + B_n. \tag{217}$$

Hence, from (198), (209), and (217), we finally obtain (7).

C Proof of Theorem 3

We can assume, without loss of generality, that the range of φ is $[0, 1]$ (otherwise, we can replace φ by $\varphi \wedge 1$). Let $\delta_k = 2^{-k}$ for all $k \ge 0$. In addition, set $\Phi = \{\varphi_k : k \ge 1\}$, where

$$\varphi_k(x) := \begin{cases} \varphi(x/\delta_k), & x \ge 0, \\ \varphi(x/\delta_{k-1}), & x < 0 \end{cases}. \tag{218}$$

Now, for any $\delta \in (0, 1]$, there exists k such that $\delta \in (\delta_k, \delta_{k-1}]$. Hence, if $f(X_i) \ge 0$, it holds that $f(X_i)/\delta_k \ge f(X_i)/\delta$, so we have

$$\varphi_k(f(X_i)) = \varphi\biggl(\frac{f(X_i)}{\delta_k}\biggr) \tag{219}$$

$$\le \varphi\biggl(\frac{f(X_i)}{\delta}\biggr), \tag{220}$$

where (220) follows from the fact that $\varphi(\cdot)$ is non-increasing.

On the other hand, if $f(X_i) < 0$, then $f(X_i)/\delta_{k-1} \ge f(X_i)/\delta$. Hence, we have

$$\varphi_k(f(X_i)) = \varphi\biggl(\frac{f(X_i)}{\delta_{k-1}}\biggr) \tag{221}$$

$$\le \varphi\biggl(\frac{f(X_i)}{\delta}\biggr), \tag{222}$$

where (220) follows from the fact that $\varphi(\cdot)$ is non-increasing.

From (220) and (222), we have

$$
\begin{aligned}
P_n\varphi_k\big(f\big) &= \frac{1}{n}\sum_{i=1}^{n}\varphi_k(f(X_i)) && (223)\\
&\le \frac{1}{n}\sum_{i=1}^{n}\varphi\bigg(\frac{f(X_i)}{\delta}\bigg) && (224)\\
&= P_n\varphi\bigg(\frac{f}{\delta}\bigg). && (225)
\end{aligned}
$$

Moreover, we also have

$$\frac{1}{\delta_k} \le \frac{2}{\delta}, \tag{226}$$

and

$$\log k = \log\log_2 \frac{1}{\delta_k} \le \log\log_2 2\delta^{-1}. \tag{227}$$

Furthermore, observe that

$$
\begin{aligned}
L(\varphi_k) &= \sup_{x\in\mathbb{R}}\left|\frac{d\varphi_k(x)}{dx}\right| && (228)\\
&= \sup_{x\in\mathbb{R}}\left|\frac{d\varphi(x/\delta_k)}{dx}\right|\mathbf{1}\{x\ge 0\} + \left|\frac{d\varphi(x/\delta_{k-1})}{dx}\right|\mathbf{1}\{x<0\} && (229)\\
&\le \frac{L(\varphi)}{\min\{\delta_k,\delta_{k-1}\}} && (230)\\
&= \frac{L(\varphi)}{\delta_k} && (231)\\
&\le \frac{2}{\delta}L(\varphi). && (232)
\end{aligned}
$$

By combining the above facts and using Theorem 2, we obtain (8) and (9).

D Proof of Theorem 6

Let $\varepsilon > 0$ and $\delta > 0$. Define recursively

$$r_0 := 1, \qquad r_{k+1} = C\sqrt{r_k\varepsilon} \wedge 1, \qquad \gamma_k := \sqrt{\frac{\varepsilon}{r_k}} \tag{233}$$

some sufficiently large constant $C > 1$ (to be determined later) such that $\varepsilon < C^{-4}$. Denote by

$$
\begin{aligned}
\delta_0 &:= \delta, && (234)\\
\delta_k &:= \delta(1-\gamma_0-\cdots-\gamma_{k-1}), && (235)\\
\delta_{k,\frac{1}{2}} &:= \frac{1}{2}\big(\delta_k+\delta_{k+1}\big), \qquad k\ge 1. && (236)
\end{aligned}
$$

For $k \ge 0$, let φ_k be a continuous function from $\mathbb{R}$ into $[0,1]$ such that $\varphi_k(u) = 1$ for $u \le \delta_{k,\frac{1}{2}}$, $\varphi_k(u) = 0$ for $u \ge \delta_k$, and linear $\delta_{k,\frac{1}{2}} \le u \le \delta_k$. For $k \ge 1$ let φ'_k be a continuous function from $\mathbb{R}$ into $[0,1]$ such that $\varphi'_k(u) = 1$ for $u \le \delta_k$, $\varphi'_k(u) = 0$ for $u \ge \delta_{k-1,\frac{1}{2}}$, and linear for $\delta_k \le u \le \delta_{k-1,\frac{1}{2}}$.

To begin with, we prove the following lemma.

Lemma 22. *Define $\mathcal{F}_0 := \mathcal{F}$, and further recursively*

$$\mathcal{F}_{k+1} := \bigg\{f\in\mathcal{F}_k : P\{f\le\delta_{k,\frac{1}{2}}\} \le \frac{r_{k+1}}{2}\bigg\}. \tag{237}$$

For all $k \geq 1$, define

$$\mathcal{G}_k := \big\{\varphi_k \circ f : f \in \mathcal{F}_k\big\}, \qquad k \geq 0 \tag{238}$$

and

$$\mathcal{G}'_k := \big\{\varphi'_k \circ f : f \in \mathcal{F}_k\big\}, \qquad k \geq 1. \tag{239}$$

Assume that

$$\begin{aligned} E^{(k)} := &\left\{\|P_n - P\|_{\mathcal{G}_{k-1}} \leq \mathbb{E}\|P_n - P\|_{\mathcal{G}_{k-1}} + \big(K_2\sqrt{r_{k-1}\varepsilon} + K_3\varepsilon\big)\sqrt{\tau_{\min}}\right\} \\ &\cap \left\{\|P_n - P\|_{\mathcal{G}'_k} \leq \mathbb{E}\|P_n - P\|_{\mathcal{G}'_k} + \big(K_2\sqrt{r_k\varepsilon} + K_3\varepsilon\big)\sqrt{\tau_{\min}}\right\}, \qquad k \geq 1, \end{aligned} \tag{240}$$

and

$$E_N := \bigcap_{k=1}^{N} E^{(k)}, \qquad N \geq 1. \tag{241}$$

Then, it holds that

$$\mathbb{P}\big[E_N^c\big] \leq 4N \exp\bigg(-\frac{n\varepsilon^2}{2}\bigg). \tag{242}$$

Proof. The proof is based on [2, Proof of Theorem 5].

For the case $C\sqrt{\varepsilon} \geq 1$, by a simple induction argument, we have $r_k = 1$. Now, without loss of generality, we assume that $C\sqrt{\varepsilon} < 1$. For this case, we have

$$\begin{aligned} r_k &= C^{1+2^{-1}+\cdots+2^{-(k-1)}}\varepsilon^{2^{-1}+\cdots+2^{-(k-1)}} &(243)\\ &= C^{2(1-2^{-k})}\varepsilon^{1-2^{-k}} &(244)\\ &= (C\sqrt{\varepsilon})^{2(1-2^{-k})}. &(245) \end{aligned}$$

From (245), it is easy to see that $r_{k+1} < r_k$ for any $k \geq 0$.

Now, observe that

$$\begin{aligned} \sum_{i=0}^{k}\gamma_i &= C^{-1}\bigg[C\sqrt{\varepsilon} + (C\sqrt{\varepsilon})^{2^{-1}} + \cdots + (C\sqrt{\varepsilon})^{2^{-k}}\bigg] &(246)\\ &= C^{-1}\bigg[(C\sqrt{\varepsilon})^{2^{-k}} + ((C\sqrt{\varepsilon})^{2^{-k}})^2 + ((C\sqrt{\varepsilon})^{2^{-k}})^{2^2} \cdots + ((C\sqrt{\varepsilon})^{2^{-k}})^{2^k}\bigg] &(247)\\ &\leq C^{-1}\bigg[(C\sqrt{\varepsilon})^{2^{-k}} + ((C\sqrt{\varepsilon})^{2^{-k}})^2 + ((C\sqrt{\varepsilon})^{2^{-k}})^{3} \cdots + ((C\sqrt{\varepsilon})^{2^{-k}})^{k}\bigg] &(248)\\ &\leq C^{-1}\sum_{i=1}^{\infty}\big((C\sqrt{\varepsilon})^{2^{-k}}\big)^i &(249)\\ &\leq C^{-1}(C\sqrt{\varepsilon})^{2^{-k}}\big(1-(C\sqrt{\varepsilon})^{2^{-k}}\big)^{-1} \leq \frac{1}{2}, &(250) \end{aligned}$$

for $\varepsilon \leq C^{-4}, C > 2(2^{\frac{1}{4}} - 1)^{-1}$ and $k \leq \log_2\log_2\varepsilon^{-1}$, where (248) follows from $i + 1 \leq 2^i$ for all $i \geq 1$ and $C\sqrt{\varepsilon} < 1$. Hence, for small enough ε (note that our choice of $\varepsilon \leq C^{-4}$ implies $C\sqrt{\varepsilon} < 1$), we have

$$\gamma_0 + \gamma_1 + \cdots + \gamma_k \leq \frac{1}{2}, \qquad k \geq 1. \tag{251}$$

Therefore, for all $k \geq 1$, we get $\delta_k \in (\delta/2, \delta)$. Note also that below our choice of k will be such that the restriction $k \leq \log_2\log_2\varepsilon^{-1}$ for any fixed $\varepsilon > 0$ will always be fulfilled.

From the definitions of (238) and (239), for $k \geq 1$, we have

$$\sup_{g\in\mathcal{G}_k} Pg^2 \leq \sup_{f\in\mathcal{F}_k} P\{f \leq \delta_k\} \leq \sup_{f\in\mathcal{F}_k} P\{f \leq \delta_{k-1,\frac{1}{2}}\} \leq \frac{r_k}{2} \leq r_k, \tag{252}$$

and

$$\sup_{g\in\mathcal{G}'_k} Pg^2 \leq \sup_{f\in\mathcal{F}_k} P\{f \leq \delta_{k-1,\frac{1}{2}}\} \leq \frac{r_k}{2} \leq r_k. \tag{253}$$

Since $r_0 = 1$, it is easy to see that (252) and (253) trivially holds at $k = 0$.

Now, by the union bound, from (240), we have

$$\begin{aligned}\mathbb{P}\big[\big(E^{(k)}\big)^c\big] \leq \mathbb{P}\bigg[&\|P_n - P\|_{\mathcal{G}_{k-1}} > \mathbb{E}\|P_n - P\|_{\mathcal{G}_{k-1}} + K_2\sqrt{r_{k-1}\varepsilon} + K_3\varepsilon\bigg]\\ &+ \mathbb{P}\bigg[\|P_n - P\|_{\mathcal{G}'_{k-1}} > \mathbb{E}\|P_n - P\|_{\mathcal{G}'_{k-1}} + K_2\sqrt{r_{k-1}\varepsilon} + K_3\varepsilon\bigg].\end{aligned} \tag{254}$$

In addition, by similar arguments which leads to (197), we have

$$\mathbb{P}\bigg\{\|P_n - P\|_{\mathcal{G}_{k-1}} \geq \mathbb{E}\big[\|P_n - P\|_{\mathcal{G}_{k-1}}\big] + u\sqrt{\frac{\tau_{\min}}{n}}\bigg\} \leq 2\exp\big(-2u^2\big). \tag{255}$$

and

$$\mathbb{P}\bigg\{\|P_n - P\|_{\mathcal{G}'_{k-1}} \geq \mathbb{E}\big[\|P_n - P\|_{\mathcal{G}'_{k-1}}\big] + u\sqrt{\frac{\tau_{\min}}{n}}\bigg\} \leq 2\exp\big(-2u^2\big). \tag{256}$$

By replacing $u = \big(K_2\sqrt{r_{k-1}\varepsilon} + K_3\varepsilon\big)\sqrt{n}$ to (255) and (256) for $K_2 > 0$ and $K_3 > 0$, we obtain

$$\begin{aligned}&\mathbb{P}\bigg\{\|P_n - P\|_{\mathcal{G}_{k-1}} \geq \mathbb{E}\big[\|P_n - P\|_{\mathcal{G}_{k-1}}\big] + \big(K_2\sqrt{r_{k-1}\varepsilon} + K_3\varepsilon\big)\sqrt{\tau_{\min}}\bigg\}\\ &\qquad \leq 2\exp\bigg(-2n(K_2\sqrt{r_{k-1}\varepsilon} + K_3)^2\bigg)\end{aligned} \tag{257}$$

and

$$\begin{aligned}&\mathbb{P}\bigg\{\|P_n - P\|_{\mathcal{G}_{k-1}} \geq \mathbb{E}\big[\|P_n - P\|_{\mathcal{G}'_{k-1}}\big] + \big(K_2\sqrt{r_{k-1}\varepsilon} + K_3\varepsilon\big)\sqrt{\tau_{\min}}\bigg\}\\ &\qquad \leq 2\exp\bigg(-2n(K_2\sqrt{r_{k-1}\varepsilon} + K_3\varepsilon)^2\bigg).\end{aligned} \tag{258}$$

Now, since $0 < C\sqrt{\varepsilon} \leq 1$, by (245), we have

$$r_{k-1} = (C\sqrt{\varepsilon})^{2(1-2^{-(k-1)})} \geq C^2\varepsilon. \tag{259}$$

Hence, from (254), (257), (258), and (259), that

$$\begin{aligned}\mathbb{P}\big[\big(E^{(k)}\big)^c\big] &\leq 4\exp\bigg(-2n(K_2\sqrt{r_{k-1}\varepsilon} + K_3\varepsilon)^2\bigg) &(260)\\ &\leq 4\exp\bigg(-2n(K_2C + K_3)^2\varepsilon^2\bigg) &(261)\\ &\leq 4\exp\bigg(-\frac{n\varepsilon^2}{2}\bigg), \qquad \forall k \geq 1 &(262)\end{aligned}$$

if we choose K_2 and K_3 such that

$$K_2C + K_3 \geq \frac{1}{2}. \tag{263}$$

Then, by the union bound and (262), we have

$$\mathbb{P}\big[E_N^c\big] \leq 4N\exp\bigg(-\frac{n\varepsilon^2}{2}\bigg). \tag{264}$$

□

Lemma 23. *Let $\varepsilon > 0$ and $0 < \alpha < 2$ such that*

$$\varepsilon \geq \left(\frac{1}{n\delta^{\alpha}}\right)^{\frac{2}{2+\alpha}} \vee \sqrt{\frac{2\log n}{n}} \vee B_n \tag{265}$$

for all large enough n, where B_n is defined in (5)*. Denote by $\mathcal{L} := \big\{f \in \mathcal{F} : P_n\{f \leq \delta\} \leq \varepsilon\big\}$. Then, on the event $E_N \cap \mathcal{L}$, we have:*

- *(i)* $\sup_{f\in\mathcal{F}_k} P_n\{f \leq \delta_k\} \leq r_k, \qquad 0 \leq k \leq N$

and

- *(ii)* $\forall f \in \mathcal{L} \quad P_n\{f \leq \delta\} \leq \varepsilon \qquad \Rightarrow f \in \mathcal{F}_N$

for all positive integer N satisfying

$$N \leq \frac{1}{\eta}\log_2\log_2 \varepsilon^{-1} \qquad \textit{and} \qquad r_N \geq \varepsilon, \tag{266}$$

and η is some implicit positive constant.

Proof. We will use induction with respect to N. For $N = 0$, the statement is obvious. Suppose it holds for some $N \geq 0$, such that $N+1$ still satisfies condition (266) of the lemma. Then, on the event $E_N \cap \mathcal{L}$ we have

$$(i) \qquad \sup_{f\in\mathcal{F}_k} P_n\{f \leq \delta_k\} \leq r_k, \qquad 0 \leq k \leq N \tag{267}$$

and

$$(ii) \qquad \forall f \in \mathcal{F} \quad P_n\{f \leq \delta\} \leq \varepsilon \quad \Rightarrow f \in \mathcal{F}_N. \tag{268}$$

Suppose now that $f \in \mathcal{F}$ is such that $P_n\{f \leq \delta\} \leq \varepsilon$. By the induction assumptions, on the event E_N defined in (241), we have $f \in \mathcal{F}_N$. Because of this, we obtain on the event E_{N+1}

$$\begin{aligned}
P\{f \leq \delta_{N,\frac{1}{2}}\} &= P_n\{f \leq \delta_{N,\frac{1}{2}}\} + (P-P_n)\{f \leq \delta_{N,\frac{1}{2}}\} && (269)\\
&\leq P_n\{f \leq \delta_N\} + (P-P_n)\{f \leq \delta_{N,\frac{1}{2}}\} && (270)\\
&\leq P_n\{f \leq \delta_N\} + (P-P_n)(\varphi_N(f)) && (271)\\
&\leq P_n\{f \leq \delta_N\} + \|P-P_n\|_{\mathcal{G}_N} && (272)\\
&\leq \varepsilon + \mathbb{E}\|P_n - P\|_{\mathcal{G}_N} + \big(K_2\sqrt{r_N\varepsilon} + K_3\varepsilon\big)\sqrt{\tau_{\min}}. && (273)
\end{aligned}$$

For the class $\mathcal{G}_N$, define

$$\hat{R}_n(\mathcal{G}_N) := \left\| n^{-1}\sum_{i=1}^{n} \varepsilon_i \delta_{X_i} \right\|_{\mathcal{G}_N}, \tag{274}$$

where ε_i is a sequence of i.i.d. Rademacher random variables. By Lemma 1, it holds that

$$\begin{aligned}
\mathbb{E}\big[\|P_n - P\|_{\mathcal{G}_N}\big] &\leq 2\mathbb{E}\big[\|P_n^0\|_{\mathcal{G}_N}\big] + B_n && (275)\\
&= 2\mathbb{E}\big[\hat{R}_n(\mathcal{G}_N)\big] + B_n. && (276)
\end{aligned}$$

From (276), we have

$$\begin{aligned}
&\mathbb{E}\big[\|P_n - P\|_{\mathcal{G}_N}\big]\\
&\quad\leq 2\mathbb{E}\big[\hat{R}_n(\mathcal{G}_N)\big] + B_n && (277)\\
&\quad= 2\mathbb{E}\big[\mathbf{1}\{E_N\}\mathbb{E}_\varepsilon\big[\hat{R}_n(\mathcal{G}_N)\big]\big] + 2\mathbb{E}\big[\mathbf{1}\{E_N^c\}\mathbb{E}_\varepsilon\big[\hat{R}_n(\mathcal{G}_N)\big]\big] + B_n. && (278)
\end{aligned}$$

Next, by the well-known entropy inequalities for subgaussian process [12], we have

$$\begin{aligned}
\mathbb{E}_\varepsilon\big[\hat{R}_n(\mathcal{G}_N)\big] \leq \inf_{g\in\mathcal{G}_N} \mathbb{E}_\varepsilon\left| n^{-1}\sum_{j=1}^{n}\varepsilon_j g(X_j)\right| \\
+ \frac{\eta}{\sqrt{n}}\int_0^{(2\sup_{g\in\mathcal{G}_N} P_n g^2)^{1/2}} H^{1/2}_{d_{P_n,2}}(\mathcal{G}_N; u)du && (279)
\end{aligned}$$

for some implicit positive constant $\eta > 0$.

By the induction assumption, on the event $E_N \cap \mathcal{L}$,

$$\begin{aligned}
\inf_{g \in \mathcal{G}_N} \mathbb{E}_\varepsilon \left| n^{-1} \sum_{j=1}^{n} \varepsilon_j g(X_j) \right| &\leq \inf_{g \in \mathcal{G}_N} \sqrt{\mathbb{E}_\varepsilon \left| n^{-1} \sum_{j=1}^{n} \varepsilon_j g(X_j) \right|^2} && (280)\\
&\leq \frac{1}{\sqrt{n}} \inf_{g \in \mathcal{G}_N} \sqrt{P_n g^2} && (281)\\
&\leq \frac{1}{\sqrt{n}} \inf_{f \in \mathcal{F}_N} \sqrt{P_n \{f \leq \delta_N\}} && (282)\\
&\leq \sqrt{\frac{\varepsilon}{n}} && (283)\\
&\leq \varepsilon, && (284)
\end{aligned}$$

where (283) follows from $\inf_{f \in \mathcal{F}_N} P_n\{f \leq \delta_N\} \leq \inf_{f \in \mathcal{F}_N} P_n\{f \leq \delta\} \leq P_n\{f \leq \delta\} \leq \varepsilon$ by the induction assumption with $f \in \mathcal{F}_N$.

We also have on the event $E_N \cap \mathcal{L}$, by (252), it holds that

$$\sup_{g \in \mathcal{G}_N} P_n g^2 \leq \sup_{f \in \mathcal{F}_N} P_n\{f \leq \delta_N\} \leq r_N. \tag{285}$$

The Lipschitz norm of φ_{k-1} and φ'_k is bounded by

$$L = 2(\delta_{k-1} - \delta_k)^{-1} = 2\delta^{-1}\gamma_{k-1}^{-1} = \frac{2}{\delta}\sqrt{\frac{r_{k-1}}{\varepsilon}} \tag{286}$$

which implies the following bound on the distance:

$$\begin{aligned}
d^2_{P_n,2}(\varphi_N \circ f; \varphi_N \circ g) &= n^{-1} \sum_{j=1}^{n} \left| \varphi_N(f(X_j)) - \varphi_N(g(X_j)) \right|^2 && (287)\\
&\leq \left(\frac{2}{\delta} \sqrt{\frac{r_N}{\varepsilon}} \right)^2 d^2_{P_n,2}(f, g). && (288)
\end{aligned}$$

Therefore, on the event $\mathcal{E}_N \cap \mathcal{L}$,

$$\begin{aligned}
&\frac{1}{\sqrt{n}} \int_0^{(2 \sup_{g \in \mathcal{G}_N} P_n g^2)^{1/2}} H^{1/2}_{d_{P_n},2}\big(\mathcal{G}_N; u\big) du\\
&\quad\leq \frac{1}{\sqrt{n}} \int_0^{(2r_N)^{1/2}} H^{1/2}_{d_{P_n},2}\left(\mathcal{F}; \frac{\delta\sqrt{\varepsilon}u}{2\sqrt{r_N}} \right) du && (289)\\
&\quad\leq \left(\frac{2\sqrt{D}}{1 - \alpha/2} \right) \left(\frac{r_N}{\varepsilon} \right)^{\alpha/4} \frac{r_N^{1/2-\alpha/4}}{\sqrt{n}\delta^{\alpha/2}} && (290)\\
&\quad\leq \left(\frac{2^{1/2+\alpha/4}\sqrt{D}}{1 - \alpha/2} \right) \frac{r_N^{1/2}}{\varepsilon^{\alpha/4}} \varepsilon^{\frac{2+\alpha}{4}} && (291)\\
&\quad= \left(\frac{2^{1/2+\alpha/4}\sqrt{D}}{1 - \alpha/2} \right) \sqrt{r_N \varepsilon}, && (292)
\end{aligned}$$

where (292) follows from the condition (265), which implies that

$$\frac{1}{n^{1/2}\delta^{\alpha/2}} \leq \varepsilon^{\frac{2+\alpha}{4}}. \tag{293}$$

From (279) and (292), we obtain that on the event $E_N \cap \mathcal{L}$,

$$\mathbb{E}_\varepsilon\big[\hat{R}_n(\mathcal{G}_N)\big] \leq \varepsilon + \frac{2^{1/2+\alpha/4}\eta\sqrt{D}}{1 - \alpha/2} \sqrt{r_N \varepsilon}. \tag{294}$$

On the other hand, we also have

$$\mathbb{E}_\varepsilon\big[\hat{R}_n(\mathcal{G}_N)\big] \le 1. \tag{295}$$

Hence, by combining with (294) and (295), from (278), we obtain

$$\begin{aligned}\mathbb{E}\big[(\big\|P_n - P\big\|_{\mathcal{G}_N})\big] &= 2\mathbb{E}\big[\mathbf{1}\{E_N\}\mathbb{E}_\varepsilon\big[\hat{R}_n(\mathcal{G}_N)\big]\big] + 2\mathbb{E}\big[\mathbf{1}\{E_N^c\}\mathbb{E}_\varepsilon\big[\hat{R}_n(\mathcal{G}_N)\big]\big] + B_n &(296)\\ &\le 2\bigg[\varepsilon + \eta\bigg(\frac{2^{1/2+\alpha/4}\eta\sqrt{D}}{1-\alpha/2}\bigg)\sqrt{r_N\varepsilon}\bigg] + 8N\exp\bigg(-\frac{n\varepsilon^2}{2}\bigg) + B_n. &(297)\end{aligned}$$

Now, by the condition (265), it holds that

$$\varepsilon \ge B_n, \tag{298}$$

and

$$\begin{aligned}8\eta N\exp\bigg(-\frac{n\varepsilon^2}{2}\bigg) &\le \frac{8\eta N}{n} &(299)\\ &\le \frac{8\log_2\log_2\varepsilon^{-1}}{n} &(300)\\ &\le \frac{8\log_2\log_2\sqrt{\frac{n}{2\log n}}}{n} &(301)\\ &\le \eta\sqrt{\frac{2\log n}{n}} &(302)\\ &\le \eta\varepsilon, &(303)\end{aligned}$$

for n sufficiently large, where (299) and (301) follows from $\varepsilon \ge \sqrt{\frac{2\log n}{n}}$, (300) follows from (266), and (302) holds for n sufficiently large.

From (297), (298), and (303), it holds that

$$\begin{aligned}\mathbb{E}\big[\big\|P_n - P\big\|_{\mathcal{G}_N}\big] &\le 2\bigg[\varepsilon + \eta\bigg(\frac{2^{1/2+\alpha/4}\sqrt{D}}{1-\alpha/2}\bigg)\sqrt{r_N\varepsilon}\bigg] + 2\varepsilon &(304)\\ &= 4\varepsilon + 2\eta\bigg(\frac{2^{1/2+\alpha/4}\sqrt{D}}{1-\alpha/2}\bigg)\sqrt{r_N\varepsilon}. &(305)\end{aligned}$$

In addition, we have

$$r_N = (C\sqrt{\varepsilon})^{2(1-2^{-N})} \ge C^2\varepsilon, \tag{306}$$

or

$$\varepsilon \le \frac{r_N}{C^2}. \tag{307}$$

Hence, from (305) and (307), we conclude that with some constant $\tilde{\eta} > 0$,

$$\mathbb{E}\big[\big\|P_n - P\big\|_{\mathcal{G}_N}\big] \le \tilde{\eta}\sqrt{r_N\varepsilon}. \tag{308}$$

From (273) and (308), on the event $E_{N+1} \cap \mathcal{L}$, we have

$$\begin{aligned}P\{f \le \delta_{N,\frac{1}{2}}\} &\le \varepsilon + \tilde{\eta}\sqrt{r_N\varepsilon} + \big(K_2\sqrt{r_N\varepsilon} + K_3\varepsilon\big)\sqrt{\tau_{\min}} &(309)\\ &\le \frac{1}{2}C\sqrt{r_N\varepsilon} &(310)\\ &= r_{N+1}/2, &(311)\end{aligned}$$

by a proper choice of the constant $C > 0$, where (311) follows from (307). This means that $f \in \mathcal{F}_{N+1}$ and the induction step for (ii) is proved.

Now, we prove (i). We have on the event E_{N+1},

$$\begin{aligned}\sup_{f\in\mathcal{F}_{N+1}} P_n\{f\le\delta_{N+1}\} &\le \sup_{f\in\mathcal{F}_{N+1}} P\{f\le\delta_{N+1}\} + \sup_{f\in\mathcal{F}_{N+1}} (P_n-P)\{f\le\delta_{N+1}\} &(312)\\ &\le \sup_{f\in\mathcal{F}_{N+1}} P\{f\le\delta_{N+1}\} + \|P_n-P\|_{\mathcal{G}'_{N+1}} &(313)\\ &\le r_{N+1}/2 + \mathbb{E}\|P_n-P\|_{\mathcal{G}'_{N+1}} + \big(K_2\sqrt{r_{N+1}\varepsilon}+K_3\varepsilon\big)\sqrt{\tau_{\min}}. &(314)\end{aligned}$$

By Lemma 1, we have

$$\begin{aligned}&\mathbb{E}\big\|P_n-P\big\|_{\mathcal{G}'_{N+1}}\\ &\quad\le 2\mathbb{E}\big[\mathbf{1}\{E_N\}\mathbb{E}_\varepsilon\big[\hat{R}_n(\mathcal{G}'_{N+1})\big]\big] + 2\mathbb{E}\big[\mathbf{1}\{E_N^c\}\mathbb{E}_\varepsilon\big[\hat{R}_n(\mathcal{G}'_{N+1})\big]\big] + B_n &(315)\\ &\quad\le 2\mathbb{E}\big[\mathbf{1}\{E_N\}\mathbb{E}_\varepsilon\big[\hat{R}_n(\mathcal{G}'_{N+1})\big]\big] + 2\mathbb{E}\big[\mathbf{1}\{E_N^c\}\mathbb{E}_\varepsilon\big[\hat{R}_n(\mathcal{G}'_{N+1})\big]\big] + \varepsilon, &(316)\end{aligned}$$

where (316) follows from the condition (265).

As above, we have

$$\begin{aligned}\mathbb{E}_\varepsilon\big[\hat{R}_n(\mathcal{G}'_{N+1})\big] &\le \inf_{g\in\mathcal{G}'_{N+1}} \mathbb{E}_\varepsilon\Big|n^{-1}\sum_{j=1}^n \varepsilon_j g(X_j)\Big|\\ &\qquad + \frac{\eta}{\sqrt{n}}\int_0^{(2\sup_{g\in\mathcal{G}'_{N+1}} P_n g^2)^{1/2}} H^{1/2}_{d_{P_n,2}}\big(\mathcal{G}'_{N+1};u\big)du. &(317)\end{aligned}$$

Since we already proved (i), it implies that on the event $E_{N+1}\cap\mathcal{L}$,

$$\begin{aligned}\inf_{g\in\mathcal{G}'_{N+1}} \mathbb{E}_\varepsilon\Big|n^{-1}\sum_{j=1}^n \varepsilon_j g(X_j)\Big| &\le \inf_{g\in\mathcal{G}'_{N+1}} \sqrt{\mathbb{E}_\varepsilon\Big|n^{-1}\sum_{j=1}^n \varepsilon_j g(X_j)\Big|^2} &(318)\\ &\le \frac{1}{\sqrt{n}}\inf_{g\in\mathcal{G}'_{N+1}} \sqrt{P_n g^2} &(319)\\ &\le \frac{1}{\sqrt{n}}\inf_{f\in\mathcal{F}_{N+1}} \sqrt{P_n\{f\le\delta_{N,1/2}\}} &(320)\\ &\le \frac{1}{\sqrt{n}}\inf_{f\in\mathcal{F}_{N}} \sqrt{P_n\{f\le\delta_{N,1/2}\}} &(321)\\ &\le \sqrt{\frac{\varepsilon}{n}} \le \varepsilon. &(322)\end{aligned}$$

By the induction assumption, we also have on the event $E_{N+1}\cap\mathcal{L}$,

$$\sup_{g\in\mathcal{G}'_{N+1}} P_n g^2 \le \sup_{f\in\mathcal{F}_{N+1}} P_n\{f\le\delta_{N,1/2}\} \le \frac{r_{N+1}}{2} \le r_N. \tag{323}$$

The bound for the Lipschitz norm of φ'_k gives the following bound on the distance

$$\begin{aligned}d^2_{P_n,2}\big(\varphi'_{N+1}\circ f;\varphi'_{N+1}\circ f\big) &= n^{-1}\sum_{j=1}^n \big|\varphi'_{N+1}\circ f(X_j) - \varphi'_{N+1}\circ g(X_j)\big|^2 &(324)\\ &\le \left(\frac{2}{\delta}\sqrt{\frac{r_N}{\varepsilon}}\right)^2 d^2_{P_n,2}(f,g). &(325)\end{aligned}$$

Therefore, on the event $E_{N+1}\cap\mathcal{L}$, we get quite similarly to (292),

$$\begin{aligned}&\frac{1}{\sqrt{n}}\int_0^{(2\sup_{g\in\mathcal{G}'_{N+1}} P_n g^2)^{1/2}} H^{1/2}_{d_{P_n,2}}\big(\mathcal{G}'_{N+1};u\big)du\\ &\qquad\le \frac{1}{\sqrt{n}}\int_0^{(2r_N)^{1/2}} H^{1/2}_{d_{P_n,2}}\left(\mathcal{F};\frac{\delta\sqrt{\varepsilon}u}{2\sqrt{r_N}}\right)du &(326)\\ &\qquad\le \left(\frac{2^{1/2+\alpha/4}\sqrt{D}}{1-\alpha/2}\right)\left(\frac{r_N}{\varepsilon}\right)^{\alpha/4}\frac{r_N^{1/2-\frac{\alpha}{4}}}{\sqrt{n}\delta^{\alpha/2}} &(327)\\ &\qquad= \left(\frac{2^{1/2+\alpha/4}\sqrt{D}}{1-\alpha/2}\right)\sqrt{r_N\varepsilon}. &(328)\end{aligned}$$

We collect all bounds to see that on the event $\mathcal{E}_{N+1} \cap \mathcal{L}$,

$$\sup_{f \in \mathcal{F}_{N+1}} P_n\{f \leq \delta_{N+1}\} \leq \frac{r_{N+1}}{2} + \bar{\eta}\sqrt{r_N \varepsilon} \tag{329}$$

for some constant $\bar{\eta} > 0$.

Therefore, it follows that with a proper choice of constant $C > 0$ in the recurrence relationship defining the sequence $\{r_k\}$, we have on the event $E_{N+1} \cap \mathcal{L}$

$$\sup_{f \in \mathcal{F}_{N+1}} P_n\{f \leq \delta_{N+1}\} \leq C\sqrt{r_N \varepsilon} = r_{N+1}, \tag{330}$$

which proves the induction step for (i) and, therefore, the lemma is proved. Finally, using Lemma 24 and the same arguments as [2, Proof of Theorem 5], we can prove Theorem 6. □

Lemma 24. *Suppose that for some $\alpha \in (0,2)$ and for some $D > 0$ such that the condition* (19) *holds. Then for any constant $\xi > C^2$, for all $\delta \geq 0$ and*

$$\varepsilon \geq \left(\frac{1}{n\delta^\alpha}\right)^{\frac{2}{2+\alpha}} \vee \sqrt{\frac{2\log n}{n}} \vee B_n, \tag{331}$$

and for all large enough n, the following:

$$\mathbb{P}\left[\exists f \in \mathcal{F} : P_n\{f \leq \delta\} \leq \varepsilon \quad \textit{and} \quad P\left\{f \leq \frac{\delta}{2}\right\} \geq \xi\varepsilon\right] \leq 4/\eta \log_2 \log_2 \varepsilon^{-1} \exp\left\{-\frac{n\varepsilon^2}{2}\right\} \tag{332}$$

and

$$\mathbb{P}\left[\exists f \in \mathcal{F} : P\{f \leq \delta\} \leq \varepsilon \quad \textit{and} \quad P_n\left\{f \leq \frac{\delta}{2}\right\} \geq \xi\varepsilon\right] \leq 4/\eta \log_2 \log_2 \varepsilon^{-1} \exp\left\{-\frac{n\varepsilon^2}{2}\right\}, \tag{333}$$

where η is some constant.

Proof. Observe that

$$\begin{aligned}
&\mathbb{P}\left[\exists f \in \mathcal{F} : P_n\{f \leq \delta\} \leq \varepsilon \wedge P\{f \leq \delta/2\} \geq \xi\varepsilon\right] \\
&\quad \leq \mathbb{P}\left[\left\{\exists f \in \mathcal{F} : \{P_n\{f \leq \delta\} \leq \varepsilon\} \wedge \{P\{f \leq \delta/2\} \geq \xi\varepsilon\}\right\} \cap E_N\right] + \mathbb{P}[E_N^c] && (334) \\
&\quad \leq \mathbb{P}\left[\left\{\exists f \in \mathcal{F}_N \cap \mathcal{L}\} \wedge \{P\{f \leq \delta/2\} \geq \xi\varepsilon\}\right\} \cap E_N\right] + \mathbb{P}[E_N^c] && (335) \\
&\quad \leq \mathbb{P}\left[\left\{\exists f \in \mathcal{F}_N \cap L\} \wedge \{P\{f \leq \delta_N\} \geq \xi\varepsilon\}\right\} \cap E_N\right] + \mathbb{P}[E_N^c] && (336) \\
&\quad \leq \mathbb{P}\left[\left\{\exists f \in \mathcal{F}_N \cap L\} \wedge \{P\{f \leq \delta_N\} > r_N\}\right\} \cap E_N\right] + \mathbb{P}[E_N^c] && (337) \\
&\quad = \mathbb{P}[E_N^c] && (338) \\
&\quad \leq 4N \exp\left(-\frac{n\varepsilon^2}{2}\right) && (339) \\
&\quad \leq 4/\eta\big(\log_2 \log_2 \varepsilon^{-1}\big) \exp\left(-\frac{n\varepsilon^2}{2}\right), && (340)
\end{aligned}$$

where (335) follows from (ii) in Lemma 23, (337) follows from $r_N \leq (C\sqrt{\varepsilon})^2 < \xi\varepsilon$ for some constant $\xi > C^2$, and (338) follows from (i) in Lemma 23, and (340) follows from the condition (266) in Lemma 23, which holds for n sufficiently large. □

Now, we return to prove Theorem 6.

Proof of Theorem 6. . Consider sequences $\delta_j := 2^{-j\frac{2}{\gamma}}$, and

$$\varepsilon_j := \left(\frac{1}{n\delta_j^{\alpha'}}\right)^{\frac{1}{2+\alpha'}}, \qquad j \geq 0, \tag{341}$$

where

$$\alpha' := \frac{2\gamma}{2-\gamma} \geq \alpha. \tag{342}$$

Then, we have

$$\varepsilon_j = n^{(\gamma-2)/4} 2^j. \tag{343}$$

Let

$$\mathcal{E} := \{\exists j \geq 0 \quad \exists f \in \mathcal{F} : P_n\{f \leq \delta_j\} \wedge P\{f \leq \delta_j/2\} \geq \xi\varepsilon_j\}. \tag{344}$$

By Lemma 24, the condition (19) implies that there exists $\xi > 0$ such that

$$\mathbb{P}[\mathcal{E}] \leq 4/\eta \sum_{j=0}^{\infty} \big(\log_2\log_2 \varepsilon_j^{-1}\big) \exp\left(-\frac{n\varepsilon_j^2}{2}\right) \tag{345}$$

$$\leq 4\upsilon' \sum_{j=0}^{\infty} \big(\log_2\log_2 n\big) \sum_{j\geq 0} \exp\left[-\frac{n^{\frac{\gamma}{2}}}{2}2^{2j}\right] \tag{346}$$

$$\leq \upsilon\big(\log_2\log_2 n\big)\exp\left[-\frac{n^{\frac{\gamma}{2}}}{2}\right] \tag{347}$$

for some $\upsilon, \upsilon' > 0$. Now, since $\hat{\delta}_n(\gamma; f) \in (0,1]$, there exists some $j \geq 1$ such that

$$\hat{\delta}_n(\gamma; f) \in (\delta_j, \delta_{j-1}]. \tag{348}$$

Then, by the definition of $\hat{\delta}_n(\gamma; f)$ in (18), we have

$$P_n\{f \leq \delta_j\} \leq P_n\{f \leq \hat{\delta}_n(\gamma; f)\} \tag{349}$$

$$\leq \sqrt{\delta_j^{-\gamma} n^{-1+\frac{\gamma}{2}}} \tag{350}$$

$$= \varepsilon_j. \tag{351}$$

Suppose that for some $f \in \mathcal{F}$, the inequality $\zeta^{-1}\hat{\delta}_n(\gamma; f) \leq \delta_n(\gamma; f)$ fails, which leads to

$$\delta_n(\gamma; f) < \zeta^{-1}\hat{\delta}_n(\gamma; f) \tag{352}$$

$$\leq \frac{\delta_{j-1}}{\zeta}. \tag{353}$$

Then, if $\zeta > 2^{1+\frac{2}{\gamma}}$, from the definition of $\delta_n(\gamma; f)$ in (17), it holds that

$$P\{f \leq \delta_j/2\} \geq P\left\{f \leq \frac{\delta_{j-1}}{\zeta}\right\} \tag{354}$$

$$\geq \sqrt{\left(\frac{\delta_{j-1}}{\zeta}\right)^{-\gamma} n^{-1+\gamma/2}} \tag{355}$$

$$= \sqrt{\frac{1}{2}2^{2j}\zeta^{\gamma} n^{-1+\frac{\gamma}{4}}} \tag{356}$$

$$= \varepsilon_j\sqrt{\frac{\zeta^{\gamma}}{4}} \tag{357}$$

$$> \xi\varepsilon_j \tag{358}$$

by choosing ζ sufficiently large, where ξ is defined in Lemma 24. From (358), it holds that $\zeta^{-1}\hat{\delta}_n(\gamma; f) \leq \delta_n(\gamma; f)$ fails for some $f \in \mathcal{F}$ to hold with probability at most $\mathbb{P}[\mathcal{E}]$.

Similarly, we can show that the event $\delta_n(\gamma; f) \leq \zeta\hat{\delta}_n(\gamma; f)$ fails for some $f \in \mathcal{F}$ with probability at most $\mathbb{P}[\mathcal{E}]$.

By using the union bound, we finally obtain (20). □

E Proof of Lemma 7

Proof. Observe that

$$\begin{aligned}&\big|P_n(f \le y) - P(f \le y)\big| \\ &\qquad = \frac{1}{n}\sum_{i=1}^{n}\big(\mathbf{1}\{f(X_i) \le y\} - \mathbb{P}(f(X_i) \le y)\big) + \frac{1}{n}\sum_{i=1}^{n}\mathbb{P}(f(X_i) \le y) - P(f \le y). \quad (359)\end{aligned}$$

Now, let

$$f_n(x) := \mathbf{1}\{f(x) \le y\} - \mathbb{P}(f(X_n) \le y), \tag{360}$$

for all $x \in \mathcal{S}$ and $n \in \mathbb{Z}^+$. It is clear that

$$\|f_n\|_\infty \le 1. \tag{361}$$

Now, let

$$g(\mathbf{x}) := \frac{1}{n}\sum_{i=1}^{n} f_i(x_i). \tag{362}$$

Then, we have

$$\begin{aligned}\big|g(\mathbf{x}) - g(\mathbf{y})\big| &= \frac{1}{n}\left|\sum_{i=1}^{n}\big(f_i(x_i) - f_i(y_i)\big)\right| &(363)\\ &\le \frac{1}{n}\left|\sum_{i=1}^{n}\big(\mathbf{1}\{f(x_i) \le y\} - \mathbf{1}\{f(y_i) \le y\}\big)\right| &(364)\\ &\le \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\{x_i \ne y_i\}. &(365)\end{aligned}$$

Then, by applying Lemma 21, it holds that

$$\mathbb{P}\left[\left|\frac{1}{n}\sum_{i=1}^{n} f_i(X_i)\right| \ge t\sqrt{\frac{\tau_{\min}}{n}}\right] \le 2\exp\big(-2t^2\big). \tag{366}$$

On the other hand, for all $y \in \mathbb{R}$, let

$$\tilde{f}_y(x) := \mathbf{1}\{f(x) \le y\}. \tag{367}$$

It is clear that $\sup_{y\in\mathbb{R}}\|\tilde{f}_y\|_\infty = 1$. Hence, by Lemma 1, it holds that

$$\begin{aligned}\left|\frac{1}{n}\sum_{i=1}^{n}\mathbb{P}(f(X_i) \le y) - P(f \le y)\right| &= \left|\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\tilde{f}_y(X_i)] - \mathbb{E}_\pi\big[\tilde{f}_y(X)\big]\right| &(368)\\ &\le \sqrt{B_n}, &(369)\end{aligned}$$

where (369) follows from Lemma 20 (with $M = 1$) and the Cauchy–Schwarz inequality.

From (359), (366), and (369), we have

$$\sup_{f\in\mathcal{F}}\sup_{y\in\mathbb{R}}\big|P_n(f \le y) - P(f \le y)\big| \le \sqrt{B_n} + t\sqrt{\frac{\tau_{\min}}{n}} \tag{370}$$

with probability at least $1 - 2\exp(-2t^2)$. □

F Proof of Lemma 9

Let $\delta > 0$. Let $\varphi(x)$ be equal to 1 for $x \le 0$, 0 for $x \ge 1$ and linear in between. Observe that

$$\begin{aligned}F_f(y) &= P\{f \le y\} &(371)\\ &\le P\varphi\left(\frac{f-y}{\delta}\right) &(372)\\ &\le P_n\varphi\left(\frac{f-y}{\delta}\right) + \big\|P_n - P\big\|_{\tilde{\mathcal{G}}_\varphi} &(373)\\ &\le F_{n,f}(y+\delta) + \big\|P_n - P\big\|_{\tilde{\mathcal{G}}_\varphi}, &(374)\end{aligned}$$

and

$$\begin{aligned} F_{n,f}(y) &\le P_n\{f \le y\} && (375)\\ &\le P_n\varphi\bigg(\frac{f-y}{\delta}\bigg) && (376)\\ &\le P\varphi\bigg(\frac{f-y}{\delta}\bigg) + \big\|P_n - P\big\|_{\tilde{\mathcal{G}}_\varphi} && (377)\\ &\le F_f(y+\delta) + \big\|P_n - P\big\|_{\tilde{\mathcal{G}}_\varphi}. && (378) \end{aligned}$$

Now, by applying Lemma 21 (see (197)), we have

$$\mathbb{P}\bigg[\|P_n - P\|_{\tilde{\mathcal{G}}_\varphi} \ge \mathbb{E}\big[\|P_n - P\|_{\tilde{\mathcal{G}}_\varphi}\big] + t\sqrt{\frac{\tau_{\min}}{n}}\bigg] \le 2\exp\big(-2t^2\big). \tag{379}$$

From (379), with probability at least $1 - 2\exp(-2t^2)$, it holds that

$$\|P_n - P\|_{\tilde{\mathcal{G}}_\varphi} \le \mathbb{E}\big[\|P_n - P\|_{\tilde{\mathcal{G}}_\varphi}\big] + t\sqrt{\frac{\tau_{\min}}{n}}. \tag{380}$$

On the other hand, from Lemma 1, we have

$$\mathbb{E}\big[\big\|P_n - P\big\|_{\tilde{\mathcal{G}}_\varphi}\big] \le 2\mathbb{E}\big[\|P_n^0\|_{\tilde{\mathcal{G}}_\varphi}\big] + B_n. \tag{381}$$

From (380) and (381), with probability at least $1 - 2\exp(-2t^2)$, it holds that

$$\|P_n - P\|_{\tilde{\mathcal{G}}_\varphi} \le 2\mathbb{E}\big[\|P_n^0\|_{\tilde{\mathcal{G}}_\varphi}\big] + B_n + t\sqrt{\frac{\tau_{\min}}{n}}. \tag{382}$$

From (374), (378), and (382), with probability at least $1 - 2\exp(-2t^2)$, we have

$$L(F_f, F_{f,n}) \le \delta + 2\mathbb{E}\big[\|P_n^0\|_{\tilde{\mathcal{G}}_\varphi}\big] + B_n + t\sqrt{\frac{\tau_{\min}}{n}}. \tag{383}$$

Furthermore, by Talagrand's contraction lemma [3, 11] for the class of function $\tilde{\varphi}(x) := \varphi(x) - 1$, we have

$$\begin{aligned} \mathbb{E}\big[\|P_n^0\|_{\tilde{\mathcal{G}}_\varphi}\big] &\le 2\mathbb{E}\bigg[\sup_{f\in\mathcal{F}, y\in[-M,M]}\bigg|\sum_{i=1}^n \varepsilon_i \frac{f(X_i) - y}{\delta}\bigg|\bigg] && (384)\\ &= \frac{2}{\delta}\mathbb{E}\bigg[n^{-1}\sup_{f\in\mathcal{F}}\bigg|\sum_{i=1}^n f(X_i)\bigg|\bigg] + \frac{2M}{\delta n}\mathbb{E}\bigg|\sum_{i=1}^n \varepsilon_i\bigg| && (385)\\ &\le \frac{2}{\delta}\mathbb{E}\big[\|P_n^0\|_{\mathcal{F}}\big] + \frac{2M}{\delta\sqrt{n}}. && (386) \end{aligned}$$

Hence, by setting $\delta := \sqrt{4\mathbb{E}[\|P_n^0\|_{\mathcal{F}}] + 4M/\sqrt{n}}$, from (383) and (386), it holds with probability at least $1 - 2\exp(-2t^2)$ that

$$L(F_f, F_{f,n}) \le 4\sqrt{\mathbb{E}[\|P_n^0\|_{\mathcal{F}}] + M/\sqrt{n}} + B_n + t\sqrt{\frac{\tau_{\min}}{n}}. \tag{387}$$

G Proof of Theorem 10

Fix $M > 0$. Since $\mathcal{F}_M \in \mathrm{GC(P)}$, we have

$$\mathbb{E}\big[\|P_n - P\|_{\mathcal{F}_M}\big] \to 0 \quad \text{a.s.} \quad n \to \infty, \tag{388}$$

which, by Lemma 1 with $t = \sqrt{\log n}$,

$$\mathbb{E}\big[\big\|P_n - P\big\|_{\mathcal{F}_M}\big] \ge \frac{1}{2}\mathbb{E}\big[\|P_n^0\|_{\mathcal{F}_M}\big] - \tilde{A}_n. \tag{389}$$

By taking $n \to \infty$, from (389), we obtain

$$\mathbb{E}\big[\|P_n^0\|_{\mathcal{F}_M}\big] \to 0, \tag{390}$$

or

$$\mathbb{E}\left[\left\|n^{-1}\sum_{i=1}^{n}\varepsilon_i\delta_{X_i}\right\|\right]_{\mathcal{F}_M} \to 0, \quad \text{as} \quad n \to \infty. \tag{391}$$

Furthermore, with $t = \sqrt{\log n}$, by Lemma 9, we have

$$\begin{aligned}\mathbb{P}\Big\{\sup_{f\in\mathcal{F}_M} L(F_{n,f}, F_f) \geq 4\sqrt{\mathbb{E}[\|P_n^0\|_{\mathcal{F}}] + M/\sqrt{n}} + B_n + \log n\sqrt{\frac{\tau_{\min}}{n}}\Big\}&\\ \leq 2\exp\big(-2\log n\big) = \frac{2}{n^2}.&\end{aligned} \tag{392}$$

It follows from (392) that

$$\begin{aligned}\sum_{n=1}^{\infty}\mathbb{P}\Big\{\sup_{f\in\mathcal{F}_M} L(F_{n,f}, F_f) \geq 4\sqrt{\mathbb{E}[\|P_n^0\|_{\mathcal{F}}] + M/\sqrt{n}} + B_n + \log n\sqrt{\frac{\tau_{\min}}{n}}\Big\}&\\ \leq 2\sum_{n=1}^{\infty}\frac{1}{n^2} < \infty.&\end{aligned} \tag{393}$$

Hence, by Borel-Cantelli's lemma [13], $B_n \to 0$, and (391), we obtain

$$\sup_{f\in\mathcal{F}_M} L(F_{n,f}, F_f) \to 0, \qquad a.s.. \tag{394}$$

Since $\sup_{f\in\mathcal{F}} L(F_{n,f_M}, F_{f_M}) = \sup_{f\in\mathcal{F}_M} L(F_{n,f}, F_f)$, from (394), we have

$$\sup_{f\in\mathcal{F}} L(F_{n,f_M}, F_{f_M}) = \sup_{f\in\mathcal{F}_M} L(F_{n,f}, F_f) \to 0, \qquad a.s.. \tag{395}$$

Now, by [2], the following facts about Levy's distance holds:

$$\sup_{f\in\mathcal{F}} L(F_f, F_{f_M}) \leq \sup_{f\in\mathcal{F}} P\{|f| \geq M\} \tag{396}$$

and

$$\sup_{f\in\mathcal{F}} L(F_{n,f}, F_{n,f_M}) \leq \sup_{f\in\mathcal{F}} P_n\{|f| \geq M\}. \tag{397}$$

Now, by the condition (29), we have

$$\sup_{f\in\mathcal{F}} P\{|f| \geq M\} \to 0 \qquad a.s. \qquad M \to \infty, \tag{398}$$

so

$$\sup_{f\in\mathcal{F}} L(F_f, F_{f_M}) \to 0, \qquad a.s. \qquad M \to \infty. \tag{399}$$

To prove that

$$\lim_{M\to\infty}\limsup_{n\to\infty}\sup_{f\in\mathcal{F}} L(F_{n,f}, F_{n,f_M}) = 0, \qquad a.s., \tag{400}$$

it is enough to show that

$$\lim_{M\to\infty}\limsup_{n\to\infty}\sup_{f\in\mathcal{F}} P_n\{|f| \geq M\} = 0, \qquad a.s. \tag{401}$$

To this end, consider the function φ from $\mathbb{R}$ into $[0,1]$ that is equal to 0 for $|u| \leq M-1$, is equal to 1 for $|u| > M$ and is linear in between. We have

$$\begin{aligned}\sup_{f\in\mathcal{F}} P_n\{|f| \geq M\} &= \sup_{f\in\mathcal{F}_M} P_n\{|f| \geq M\} && (402)\\ &\leq \sup_{f\in\mathcal{F}_M} P_n\varphi(|f|) && (403)\\ &\leq \sup_{f\in\mathcal{F}_M} P\varphi(|f|) + \|P_n - P\|_{\mathcal{G}} && (404)\\ &\leq \sup_{f\in\mathcal{F}_M} P\{|f| \geq M-1\} + \|P_n - P\|_{\mathcal{G}}, && (405)\end{aligned}$$

where

$$\mathcal{G} := \big\{\varphi \circ |f| : f \in \mathcal{F}_M\big\}. \tag{406}$$

Then, by using the same arguments to obtain (197), it holds with probability $1 - 2\exp(-2t^2)$ that

$$\|P_n - P\|_{\mathcal{G}} \leq \mathbb{E}\big[\|P_n - P\|_{\mathcal{G}}\big] + t\sqrt{\frac{\tau_{\min}}{n}} \tag{407}$$

$$\leq 2\mathbb{E}[\|P_n^0\|_{\mathcal{F}}] + A_n + t\sqrt{\frac{\tau_{\min}}{n}}. \tag{408}$$

Then, by setting $t = \log n$ and using the Borel-Cantelli's lemma [13], the following holds almost surely:

$$\|P_n - P\|_{\mathcal{G}} \leq 2\mathbb{E}\big[\|P_n^0\|_{\mathcal{G}}\big] + A_n + \log n\sqrt{\frac{\tau_{\min}}{n}}. \tag{409}$$

Now, since $\varphi \circ f \in \varphi \circ \mathcal{F}_M$, by (391) and Talagrand contraction lemma [3, 12], we have

$$\mathbb{E}\big[\|P_n^0\|_{\mathcal{G}}\big] \to 0 \qquad \text{as} \quad n \to \infty. \tag{410}$$

From (409) and (410), we obtain

$$\|P_n - P\|_{\mathcal{G}} \to 0 \quad a.s.. \tag{411}$$

Hence, we obtain (ii) from (i), the condition (29), and (411).

To prove that (ii) implies (i), we use the following bound [14]

$$\bigg|\int_{-M}^{M} td(F-G)(t)\bigg| \leq \upsilon L(F,G), \tag{412}$$

which holds with some constant $\upsilon = \upsilon(M)$ for any two distribution functions on $[-M, M]$. This bound implies that

$$\|P_n - P\|_{\mathcal{F}_M} = \sup_{f \in \mathcal{F}_M} |P_n f - Pf| \tag{413}$$

$$\leq \sup_{f \in \mathcal{F}_M} \bigg|\int_{-M}^{M} td(F_{n,f} - F_f)(t)\bigg| + M \sup_{f \in \mathcal{F}_M} \big|P(|f| \geq M) - P_n(|f| \geq M)\big| \tag{414}$$

$$\leq \upsilon \sup_{f \in \mathcal{F}_M} L(F_{n,f}, F_f) + M \sup_{f \in \mathcal{F}_M} \big|P(|f| \geq M) - P_n(|f| \geq M)\big| \tag{415}$$

$$\leq \upsilon \sup_{f \in \mathcal{F}} L(F_{n,f}, F_f) + M \sup_{f \in \mathcal{F}_M} \big|P(|f| \geq M) - P_n(|f| \geq M)\big|. \tag{416}$$

Now, by Lemma 7, with probability at least $1 - 2\exp(-2t^2)$, the following holds:

$$\sup_{y \in \mathbb{R}} \sup_{f \in \mathcal{F}_M} \big|P_n(f \leq y) - P(f \leq y)\big| \leq t\sqrt{\frac{\tau_{\min}}{n}} + \sqrt{B_n}. \tag{417}$$

By setting $t = \sqrt{\log n}$ and using the Borel-Cantelli's lemma, from (417), we obtain

$$\sup_{y \in \mathbb{R}} \sup_{f \in \mathcal{F}_M} \big|P_n(f \leq y) - P(f \leq y)\big| \to 0, \quad a.s.. \tag{418}$$

Finally, from (417), (418), and (ii), we obtain (i). This concludes our proof of Theorem 10.

H Proof of Theorem 12

The proof is based on [2, Proof of Theorem 9]. Since $\mathcal{F}$ is uniformly bounded, we can choose $M > 0$ such that $\mathcal{F}_M = \mathcal{F}$. To prove the first statement, note that $\mathcal{F} \in \mathrm{BCLT(P)}$ means that

$$\mathbb{E}\big[\|P_n - P\|_{\mathcal{F}}\big] = O(n^{-1/2}). \tag{419}$$

Now, from Lemma 1, we have

$$\mathbb{E}\big[\big\|P_n - P\big\|_{\mathcal{F}}\big] \geq \frac{1}{2}\mathbb{E}\big[\|P_n^0\|_{\mathcal{F}}\big] - \tilde{A}_n, \tag{420}$$

for all $t > 0$. By applying (420), it easy to see that

$$\mathbb{E}\bigg[\bigg\|n^{-1}\sum_{i=1}^{n}\varepsilon_i\delta_{X_i}\bigg\|_{\mathcal{F}}\bigg] = \mathbb{E}\big[\|P_n^0\|_{\mathcal{F}}\big] \tag{421}$$

$$\leq 2\tilde{A}_n + 2\mathbb{E}\big[\big\|P_n - P\big\|_{\mathcal{F}}\big] \tag{422}$$

$$\leq O\bigg(\sqrt{\frac{\log n}{n}}\bigg) \tag{423}$$

since $\tilde{A}_n = O\big(\sqrt{\frac{\log n}{n}}\big)$ by (1).

Now, from Lemma 9, for $t = \sqrt{\log n}$,

$$\mathbb{P}\bigg\{\sup_{f\in\mathcal{F}} L(F_f, F_{f,n}) \geq 4\sqrt{\mathbb{E}[\|P_n^0\|_{\mathcal{F}}] + M/\sqrt{n}} + B_n + \sqrt{\frac{\tau_{\min}\log n}{n}}\bigg\} \leq \frac{2}{n^2}. \tag{424}$$

From (423) and (424), it holds that

$$\mathbb{P}\bigg\{\bigg(\frac{n}{\log n}\bigg)^{1/4}\sup_{f\in\mathcal{F}} L(F_{n,f}, F_f) \geq D\bigg\} \to 0 \tag{425}$$

as $n \to \infty$ for some constant D, or

$$\sup_{f\in\mathcal{F}} L(F_{n,f}, F_f) = O_P\bigg(\bigg(\frac{\log n}{n}\bigg)^{1/4}\bigg). \tag{426}$$

Now, recall

$$\tilde{\mathcal{G}}_\varphi := \bigg\{\varphi \circ \bigg(\frac{f - y}{\delta}\bigg) - 1 : f \in \mathcal{F}, y \in [-M, M]\bigg\}. \tag{427}$$

To prove the second statement, we use the following fact [2, p.29]:

$$\mathbb{E}_\varepsilon\big[\|P_n^0\|_{\tilde{\mathcal{G}}_\varphi}\big] \leq \frac{d}{\sqrt{n}}\bigg[\int_0^{\sqrt{2}} H_{d_{P_n,2}}^{1/2}(\mathcal{F}; \delta u)du + \sqrt{\log\frac{4M}{\delta}} + 1\bigg] \tag{428}$$

for some constant d, which, under the condition (32), satisfies

$$\mathbb{E}_\varepsilon\big[\|P_n^0\|_{\tilde{\mathcal{G}}_\varphi}\big] \leq d\bigg[\frac{1}{\delta^{\alpha/2}}\sqrt{\frac{1}{n}} + \frac{1}{\sqrt{n}}\bigg(\sqrt{\log\frac{4M}{\delta}} + 1\bigg)\bigg]. \tag{429}$$

Now, by Lemma 9, it holds for all $t > 0$ and $\delta > 0$ that

$$\mathbb{P}\bigg\{\sup_{f\in\mathcal{F}} L(F_f, F_{f,n}) \geq \delta + \mathbb{E}\big[\|P_n^0\|_{\tilde{\mathcal{G}}_\varphi}\big] + B_n + t\sqrt{\frac{\tau_{\min}}{n}}\bigg\} \leq 2\exp(-2t^2). \tag{430}$$

Since $\mathbb{E}\big\|n^{-1}\sum_{i=1}^n \varepsilon_i\delta_{X_i}\big\|_{\tilde{\mathcal{G}}_\varphi} = \mathbb{E}[\|P_n^0\|_{\tilde{\mathcal{G}}_\varphi}] \leq d\big[\sqrt{\frac{\log n}{n}}\delta^{-\alpha/2} + \frac{1}{\sqrt{n}}\big(\sqrt{\log\frac{4M}{\delta}} + 1\big)\big]$, from (430), for all $t > 0$, we have

$$\mathbb{P}\bigg\{\sup_{f\in\mathcal{F}} L(F_f, F_{f,n}) \geq \delta + d\bigg[\sqrt{\frac{\log n}{n}}\delta^{-\alpha/2} + \frac{1}{\sqrt{n}}\bigg(\sqrt{\log\frac{4M}{\delta}} + 1\bigg)\bigg] + B_n + t\sqrt{\frac{\tau_{\min}}{n}}\bigg\}$$
$$\leq 2\exp\big(-2t^2\big). \tag{431}$$

Now, by choosing $t = \sqrt{\log n}$ and $\delta = (\log n)n^{-\frac{1}{2+\alpha}}$, we have

$$\mathbb{P}\bigg\{\sup_{f\in\mathcal{F}} L(F_f, F_{f,n}) \geq \nu(\log n)n^{-\frac{1}{2+\alpha}}\bigg\} \leq \frac{2}{n^2} \tag{432}$$

for some constant ν and for $n \geq N_0$ for some finite N_0 big enough.

From (432), we have

$$\sum_{n=1}^{\infty} \mathbb{P}\bigg\{ \sup_{f \in \mathcal{F}} L(F_f, F_{f,n}) \geq \nu (\log n) n^{-\frac{1}{2+\alpha}} \bigg\} \leq N_0 + \sum_{n=N_0}^{\infty} \frac{2}{n^2} < \infty. \tag{433}$$

Hence, by Borel-Cantelli's lemma [13], it holds that

$$\sup_{f \in \mathcal{F}} L(F_f, F_{f,n}) = O_P\big((\log n) n^{-\frac{1}{2+\alpha}}\big), \qquad a.s. \tag{434}$$

This concludes our proof of Theorem 12.

References

[1] A. W. Van Der Vaart and Wellner. *Weak convergence and Empirical Processes.* Springer, New York, 1996.

[2] V. Koltchinskii and D. Panchenko. Empirical Margin Distributions and Bounding the Generalization Error of Combined Classifiers. *The Annals of Statistics*, 30(1):1 – 50, 2002.

[3] Lan V. Truong. On rademacher complexity-based generalization bounds for deep learning. *ArXiv*, abs/2208.04284, 2022.

[4] Peter Bartlett, Yoav Freund, Wee Sun Lee, and Robert E. Schapire. Boosting the margin: a new explanation for the effectiveness of voting methods. *The Annals of Statistics*, 26(5):1651 – 1686, 1998.

[5] A. G. Wilson and P. Izmailov. Bayesian deep learning and a probabilistic perspective of model construction. In *Proc. 37th ICML*. Morgan Kaufmann, 2020.

[6] Pascal Lezaud. Chernoff and Berry-Esséen inequalities for Markov processes. *ESAIM: Probability and Statistics, EDP Sciences*, 5:183–201, 2001.

[7] D. Rudolf. Explicit error bounds for Markov chain Monte Carlo. *Arxiv: 1108.3201*, 2011.

[8] Shravas Rao. A Hoeffding inequality for Markov chains. *Electronic Communications in Probability*, 24(none):1 – 11, 2019.

[9] J. Fan, B. Jiang, and Q. Sun. Hoeffding's Inequality for General Markov Chains and Its Applications to Statistical Learning. *J. Mach. Learn. Research*, 22:1–35, May 2021.

[10] Daniel Paulin. Concentration inequalities for Markov chains by Marton couplings and spectral methods. *Electronic Journal of Probability*, 20(79):1 – 32, 2015.

[11] M. Ledoux and M. Talagrand. *Probability in Banach Spaces.* Springer, New York., 1991.

[12] A. W. Van Der Vaart and J. A. Wellner. *Weak convergence and Empirical Processes.* Springer, New York., 1996.

[13] P. Billingsley. *Probability and Measure.* Wiley-Interscience, 3rd edition, 1995.

[14] I. Kontoyiannis, L. A. Lastras-Montaño, and S. P. Meyn. Relative entropy and exponential deviation bounds for general Markov chains. In *Proc. of Intl. Symp. on Inform. Th.*, 2005.